

Moral Compliance and the Concealed Charm of Prudence

by Jan Tullberg

Abstract:

The article discusses morality from the perspective of compliance. The expectation that rules easily will be internalized and then followed makes very unrealistic assumptions about human behavior. There are good reasons to put more trust in prudence - a morality supported by incentives - than pure morality. There are reasons to be critical to morals that have no prospect to be developed into prudence. Such a preference for a prudent morality is also a criterion for selection among different suggestions of moral rules.

A normative suggestion in the Prisoner's Dilemma exemplifies the case of prudent morality. The player should play contribute if the expected value of that choice exceeds his payoff of mutual defection. This suggestion is in line with Gauthier's rules for constrained maximizers. Empirical evidence indicates strong preferences for symmetric outcomes: People tend to contribute rather than defect if the other player also is expected to contribute; otherwise mutual defection is the preferred outcome. Fruitful revisions of the game are more likely to be found in modifying the payoffs of playing defection than in seeing the players' priority of their personal payoff as the main obstacle for cooperation.

The article questions the value of an absolute morality and suggests a morality that is more of a prudent policy than a categorical imperative. A conviction of a good average result in the long run is the most relevant argument for complying with a rule.

Introduction

Regardless of the kind of moral system one proposes, there is a need for some practical justification. The question 'What is right?' is closely followed by the questions 'Why should I do what is right?' and 'Why should I expect others to do what is right?' These two questions of compliance are central if one's suggestion of morals aims to be useful as social morality and not only a thought. Thomas Hobbes termed the central moral agreement 'The Second Law of Nature', but the main theme of the article is to penetrate the problem Hobbes called 'Third Law of Nature'. That is, it is not enough to reach a moral agreement - there is also a need for compliance to this moral agreement.

Since there is some interdependence between morality and compliance, moral positions are most relevant for the discussion. In this article, the chief candidate is the position that moral principles should be built upon reciprocal agreement with the purpose of enhancing individual self-interest. Such a moral position will be specified for the case of the Prisoner's Dilemma.

Self-interest is not only part of general human behavior, it is also often a reason for obeying rules. A prudent consideration of punishment is one reason for following the constraints of morality, but it is generally not considered enough. Even philosophers in the rational mould as David Gauthier (1986, 1990) ask for more. True virtue and pure morality are generally seen as forces capable of accomplishing what is out of reach for prudence. However, there are reasons to doubt the wisdom of honoring virtue at the expense of prudence.

1 Different kinds of compliance

A classic discussion centers around prudence versus morality with participants like Adam Smith and Henry Sidgwick. A central component in prudent behavior is guidance by self-interest while moral behavior is seen as determined by a strong sense of doing right. However, such a distinction does not give a clear-cut split into two different groups. In many situations self-interest and rightness are more complementary than opposites. For example, murder is a crime with a high detection rate and a severe punishment, so prudence weighs heavily against it. Still, such negative incentives can hardly be seen as opposed to a personal or general conviction that murder is wrong. Rather there is a split in two dimensions: An action can be supported or not by intrinsic reason - morality; and it can be supported or not by extrinsic reason such as reward or punishment. It might be instructive to see four combinations of these two major components:

The first category is prudent morality, labeled just 'Prudence'. There is moral agreement involved, but social punishment or reward provides strong incentives for compliance. The agent might be intrinsically motivated, but extrinsic incentives might be sufficient explanations for the agent's behavior. This morality is incentive compatible.

The second category is non-prudent morality or 'Pure Morality'. The action is not explained by prudence, since the agent acting to the moral idea often obtains more disadvantages than advantages.

In the third category - 'Pressure' - incentives are so strong that the agent complies with the sustained norms in spite of misgivings. The incentives are, however, not necessarily negative; certain dubious behavior is likely to generate money, social admiration or promotion. When these incentives outweigh the agent's own moral conviction, he complies with the social morality.

The fourth category - 'Soft Conformism' - lacks both personal conviction and substantial incentives. That the agent still complies can be seen as a behavior similar to purchases of low involvement products. With a low cost and low involvement the chosen alternative might be the one the agent picked previously or the one others choose. Behavior in the fourth group is the most likely to fail in obtaining compliance; here are the fading taboos of the past and costumes that lost their *raison d'être*.

Of course, a categorization like this one is a simplification and there are shades of gray making some specific acts hard to classify. But this structure might prove itself to be helpful for an analysis.

Figure 1

Reasons for compliance. Different motives for actions according to a rule.

	Incentive support	No incentive support
Intrinsic support	Prudence	Pure Morality
No intrinsic support	Pressure	Soft Conformism

It might be argued that the category pressure is the antithesis to pure morality and therefore the proper heir to the term prudence, but it is usually dubious supporters, not opponents, that are scrutinized and separated from pure morality. A common view is that prudence can bring moral compliance up to a certain point beyond which there might be no external monitoring or no incentive linked to that behavior. The question then arises whether the moral belief is strong enough. If you get the chance to kill an adversary with no risk of getting caught and punished, would you take it? If a man is restricted by his conscience he is always constrained, but an extrinsic control does not have the same permanent presence. Although pure morality can take the place of prudence, prudence cannot fill the space left by pure morality. At least that is the claim of the proponents of pure morality.

A choice of pure morality over prudent morality does not necessarily imply that its supporters have to pursue the difficult argumentation line of 'less is more', i.e. rules without incentives are better than these rules with incentives. If prudence is preferred this indicates a constraint upon which rules to accept as suitable. So prudent morality will imply not only more supported rules, but also fewer and less demanding rules. Prudent morality is a morality of performance rather than a morality of ambition.

2 A suggestion of morality for the Prisoner's Dilemma

After this clarification of morality and incentives it might be useful to take a common example in the discussion of morality: the Prisoner's Dilemma. As mentioned I will take a position in the normative question how one should behave in such a situation. The reasoning will be illustrated with calculations from the following matrix:

Figure 2 Prisoner's Dilemma

		Person B	
		contribute	defect
Person A	Contribute	3; 3 (Cooperation)	-1; 4 (A is a Sucker/Saint)
	Defect	4; -1 (A is free-riding)	2; 2 (Non-Cooperation)

Most often the two choices of action are called 'defect' and 'cooperate', but I think that the latter term should be used for a specific outcome. It takes two to tango, and the action should be seen as a contribution that might result in a cooperation but not necessarily so.

The game theoretical judgment is that a rational person should always choose to defect since that gives the best result regardless of the other person's choice of action (if A chooses to defect he will get 4 instead of 3, or 2 instead of -1 depending upon B's choice). Therefore the players end up in the Non-Cooperation outcome of (2; 2). If one player deviates from and chooses to contribute he will get an impairment to -1. Non-Cooperation becomes a Nash equilibrium; implying that none of the players will change his strategy. This is remarkable since both players are better off in the Cooperation solution (3;3), (Non-Cooperation is Pareto dominated by Cooperation). A problematic situation is illustrated when rationality and efficiency promote two different alternatives. Many game theorists (e.g. Binmore 1998) insist that the Nash equilibrium is the solution and when people behave differently this is caused by confusion or by the game played not really being a pure Prisoner's Dilemma. However, the social dilemmas that are of interest for most scientists are not pure

Prisoner's Dilemmas and the interesting cases are the variations when short term interest confronts reputation effects, communicative possibilities and long term considerations. The discussion in this paper of the Prisoner's Dilemma is in the latter 'social science situation', not the pure situation of Game Theory.

Let me present a normative suggestion: A should contribute if the expected value of the contribute alternative exceeds the value of non-cooperation; in this example 2. If person A makes the estimate - based upon rules in society, reputation of B, previous experience etc. - that the probability is 0.8 that B will contribute and 0.2 that he will defect, the expected value of playing contribute will be 2.2. ($0.8 \times 3 + 0.2 \times -1$). Expressed in more moralistic terms actor A excludes that he will free-ride himself, but includes in his contribute alternative realistic estimates of the other agent free-riding. This principle does not oblige A to follow a behavior that is just advantageous in a perfect world, but just to participate fairly in cooperation with profitable forecast. A is not obliged to lose in a world of defectors - becoming a person sometimes respectfully called 'a saint', but often in game theory labeled 'a sucker'. However, this suggested rule does restrict him from improving his revenue further by free-riding. The rule is in line with Gauthier's reasoning for 'constrained maximization' ('constrained' because of the inhibition to free-ride and 'maximization' because of the striving for personal utility) and that term is used synonymously with my preferred term 'reciprocal' that emphasizes other aspects.

This conditioning of one's behavior upon the expected behavior of the inter actor might cause objection from persons advocating a more categorical or universal morality, but from a contractarian point of view it sounds much more reasonable. If morals are mutual agreements, it is natural that behavior is influenced by tacit or explicit rules and more ultimately actual behavior from the people one interacts with. In line with the Darwinian theory of natural selection, human behavior can be expected to be adaptive in general, and intelligence may be viewed as an adaptation to variation and change in environments. In a contractarian and Darwinian perspective it is reasonable to see morality as being a predisposition in line with intelligence and long-term rationality, not as a countervailing force to reason and interest.

Frank (1988) presents an experiment of the Prisoner's Dilemma in which the test persons were asked about their expectations of the other person's behavior. Of those that thought the other person would contribute, 83 per cent decided to contribute. The share of contribute among persons that thought that the other person would not contribute was only 15 per cent. The same pattern has been shown in work such as Dawes et al. (1977). This experiment also shows that one explanation for this difference is that the expectations partly are rationalizations of a chosen decision. Such a behavior indicates a normative support for a symmetric outcome. People do not want to be a sucker/saint and they don't want to be a free-rider - they want to be cooperative if there is a good possibility for success, and a non-cooperator when the other person will not contribute. To get a symmetric result, they adjust their own behavior and also their expectations of the other players' behavior.

I think Thomas Hobbes would have supported this rule of reciprocal morality for the Prisoner's Dilemma, but his nemesis in *Leviathan* - the Foole - would certainly object. The Foole supported the idea that it is rational to make agreements in

situations like this game, but that it is another matter if one should honor such an agreement. Supporting the game theorists, he would point out that one square in the matrix is excluded in the suggested calculation. If the players get a chance to talk before the game - as they do in some variations - this is normally classified as "cheap talk". Of course people say they will contribute, but rational people will discard such promises and if dealing with other rational persons, a rational person will waste little time talking or listening to such cheap talk. Being more practically oriented than the game theorist, the Foole will not think his listeners are all rational persons. Even if he does not have the best of reputation, he might get a significant improvement by fooling some of the people some of the time. Fifteen per cent suckers/saints change the expected value of the defect alternative significantly: $0.85 \times 2 + 0.15 \times 4 = 2.3$. Non-cooperation is not the only outcome of playing defect. In similar manner as the reciprocal agent burdens the contribute alternative with the risks of defection from the co-player, the Foole sweetens the defect alternative with own free-riding in addition to the non-cooperative outcome. Why not? This question pinpoints the crucial difference between a reciprocal morality and straight maximization. From a practical point of view, a lot can be said for non-cooperation. Moral judgements will distinguish between a person abstaining from cooperation and a person defecting from given promises. By a declaration of not cooperating the other player can avoid being exploited.

The proponents of straight maximization are not alone in the argument against a reciprocal morality. Many supporters of more conventional morality support the Foole's critique that a rational person will break the reciprocal convention when he profitably can do so. If there is a series of iterated prisoners dilemma, it is not a problem to claim that the reciprocal can do better than the person starting with a gain by defection. The following cooperative sequence of four outcomes: "second best, second best, second best, second best", beats a straightforward sequence of: "best, third best, third best, third best", if using the payoffs of this example (12 points versus 10 points). But the critics of reciprocity point out that if the straightforward maximizer chooses his defection with care, he will always beat the reciprocal sequence. The straightforward optimal alternative is: "second best, second best, second best, best" (13 points).

Österberg writes "A straightforward maximizer should, of course, always reckon with the possibility that an act of carrying out an agreement may be known and hinder him from making new agreements. One single act of defecting may have disastrous consequences for him. So a prudent straightforward maximizer, living among people who normally keep their agreements (whether because they are constrained maximizers or prudent straightforward maximizers), also normally keeps his agreements" (Österberg 1988, p 166)

But an agent only chooses an action or a strategy, but not an outcome. Maximax as a hope does not eliminate expected value as a rational forecast. You cannot choose to be a successful bank robber; you could only try to be one. You can decide between a strategy of robbing banks by hold-ups or of nightly break-ins, but you can still not choose to be a successful bank robber. By the same token you can choose a policy of act egoism or reciprocity or altruism, but there is not any alternative of cherry picking with *ex post* knowledge in advance. The outcome of the policy is a separate story.

It might never be discovered whether a person, that behaves reciprocally, is really in his heart a maximizer deceiving others. Perhaps the person never gets an opportunity he thinks is profitable enough for defecting. Therefore, the successfully camouflaged opportunist might more often be an assimilated conformist than a hidden saboteur. An instructive example might be a reflection over opportunists following strong extrinsic incentives of moral ideas few sympathize with today, that is, the National-Socialistic Party of Germany. The Nazis themselves shared the moralists' contempt for people who joined a popular movement without a strong belief. These were named 'Sunday Nazis' by the true believers (Drucker 1939). However, this unity of opinion about opportunism does not imply that the moralists are right about the importance of intrinsic beliefs. For many victims the difference between a Nazi by virtue and a Nazi by convention might not be very significant. How much of a divergence is caused by a lack of intrinsic belief? (Browning 1992).

3 Sustaining reciprocal morality in the Prisoner's Dilemma

If the reciprocal person is unaffected by the temptation of the Foole, there are two developments of the present situation that might influence a change in behavior. One threat is that many other reciprocals change their minds and the group becomes such a low proportion of the population that our agent will change his policy to defect. At a low proportion of reciprocals in the population it might be right to discriminate against all that are not confirmed reciprocals. A second more expansive policy is to attack the profitability of free-riding.

If reciprocals could increase their share of the population to 0.9 this would generate an increase in the expected value of contribution to 2.6. One way to sustain the behavior and improve its payoff is to punish defectors. If we assume that two in three free-riders can be caught and fined -2, the defect alternative will get a reduced payoff to 1.7 ($0.85 \times 2 + 0.10 \times -2 + 0.05 \times 4 = 1.7$). With such bleak possibilities many agents will turn to contribution and our reciprocals will get a better situation, to such an extent that it will motivate paying for the sword that provides the beneficial motivation. As Hobbes puts it "Covenants being but words, and breath, have no force to oblige, contain, constrain or protect any man, but what it has from the publique Sword" (Hobbes 1651). This is not to trust pure morality, but to choose prudent morality. If the prime goal is to generate the right behavior this motivation seems adequate for a rational person, who might understand and adjust even if he remains unreformed. One objection to the public sword is that when starting to use force there is little restriction from a development from prudence to pressure. Incentives can generate desired behavior, but there is no guarantee that the behavior is morally right.

I think this is a valid point and brings up the risk of degeneration. What one person judges to be a justly rewarded behavior might be an example of the effects of the corruption of the times in another persons judgement. But there is little indicating a more regular problem of incentives such as a larger social impact for misguided rules and incentives than those pushing in the right direction. There is not much supporting the idea of a general decay, but rather that some trial and error will direct efforts towards curbing problematic asocial behavior rather than terrorizing ordinary citizens. But there are other suggestions for obtaining the cooperative solution in the

Prisoner's Dilemma. Instead of an improved support structure, 'higher' moral norms, that is less self interested concerns, are seen as the crucial factor.

4 Unselfishness as a solution to the Prisoner's Dilemma.

The Prisoner's Dilemma is often used, not as a dilemma, but as an example of how egoism brings about sub optimal results. However, this is hardly more than a result of framing. The Prisoner's Dilemma can easily be adjusted so that suspicions will indicate another villain.

If the reader is unconvinced of altruism being caught in a Prisoner's Dilemma an example might make a clear-cut case. Two believers do completely ignore their personal good - including factors such as reputation, status and salvation - and look only to the divine ranking of different outcomes. The choice for each player stands between two actions: Tolerance and Holy War. Person A follows what he thinks is Allah's preferences, and Person B follows what he think is Jehovah's preferences.

According to A, Allah has the following ranking of outcomes:

- All will convert to the right faith (Holy War; tolerance). Payoff 4.
- Sinners and faithful live in peace (Tolerance; tolerance). Payoff 3.
- Conflict between faithful and sinners (Holy War; holy war). Payoff 2.
- The true faith is trashed by the sinners (Tolerance; holy war). Payoff -1.

According to Person B, Jehovah's ranking order, outcome (Holy War; tolerance) and (Tolerance; holy war) trade places, and the *dejà vu* suspicion is confirmed.

Figure 3

		Person B	
		tolerance	holy war
Person A	Tolerance	3; 3	-1 ; 4
	Holy War	4; -1	2 ; 2

As documented in history it is also possible to believe in the same God and the same holy scripts, but still hold very different opinions about God's preferences. The central problem in the Prisoner's Dilemma is not selfish preferences, but different preferences. There seems to be more than one mental block obstructing an optimal solution. Even in our relatively materialistic world there is considerable idealistic positioning. Religious commitment, political utopias and effort to do good, are all likely candidates for generating sub optimal social outcomes.

Another suggestion that is highly regarded not only in religious circles, is to turn the other cheek and suffer, rather than give an eye for an eye. This approach suggests less focus at the own payoff and instead taking the payoff of the other player more into account. This can be motivated with a belief in the moral primacy of others which is difficult to justify. But it can also be motivated as a superior strategy to curb violence and promote social behavior by the power of the good example. For the PD-game the rule implies that one should always contribute even if the other is not.

For an analyst of games this sounds like a most dubious advise. Apart from being costly to the agent deliberately acting saintly, there is a further problem. When one agent becomes a saint, another agent becomes a free-rider. Regardless of the generous agent's intention, he is supporting asocial behavior. Even if ignoring his own cost, he will, by accepting and supporting free-riding, undermine the expected value for other agents of playing contribute. In a second shift this might influence responsible individuals that are reciprocal but not altruistic to reconsider their strategy to play contribute. Saintly behavior is often described as a step from Non-cooperation towards Cooperation, but seen in its social context it is often more likely to be a step from Cooperation towards Non-cooperation. Furthermore, is there not a lot more to say for symmetric Non-cooperation itself than for an asymmetric transfer from the giver to the taker?

Even, if making the demanding assumption of general benevolence, the saintly strategy does not imply a general perfect solution. The discussed matrix is not necessarily the only game in social life. Mandeville and Adam Smith have both made a strong case for the mechanism of private vices to public benefits. In striving for the own good you produce more social good than if you make efforts to help your fellow man. The payoff of the upper left and the lower right square might trade place.

Figure 4

		Person B	
		contribute	defect
Person A	Contribute	2 ; 2	-1 ; 4
	Defect	4 ; -1	3 ; 3

Greedy consumers will, in their effort to get value for themselves, put hard pressure upon producers for lower prices and more products. Those results, lower prices and more demand for labor, are likely to bring more benefits than selfless contributions to systems of charity.

Furthermore, it seems rather more reasonable to take into account individual preferences when they contain private information - a reason to take special account of your own preferences rather than making assumptions about others'. With

principled preferences or higher authorities there are less rationale for an amalgam of preferences. Such high level moral authority will disregard not only negative effects for the self, but such effects for 'sinners' and 'insignificant others' are likely to carry even less weight.

5 The attraction of pure morality

One line of critique is that an attraction to prudence is considered a sign of weakness in the belief regarding the capacity of maintaining restrictions by pure morality. The punishment suggested to sustain a contributing behavior in a Prisoner's Dilemma might transform it from an ethical decision to a maximizing decision. Anytime the incentives or the control over incentives is insufficient there will be a weaker defense hindering defection (Frohlich & Oppenheimer 1992). Then an instrumental treatment of morality undercuts the function of morality as an effective constraint upon behavior. As with Michael Ruse (1986) one can see morality as an illusion of objectivity, but if the illusion dissolves, morality will unravel. One way of supporting the illusion is to argue that it is the necessary and sufficient foundation for moral behavior.

Many a philosopher talks of the primacy of virtue (e.g. MacIntyre 1981). While intrinsic values are seen as strong and permanent, instrumental values come and go. However, both the order and the strength between them differ widely. A political opinion that is stronger than most is a preference for democracy. By sheer enthusiasm it is sometimes attributed intrinsic value, but that is intellectually difficult to justify. However, democracy's instrumental value is so significant that it carries more weight than many intrinsic opinions. It is hard to see a good reason to value the intrinsic as more important and reliable than the extrinsic/instrumental; it is even hard to draw a sharp line between them. From an evolutionary analysis it is reasonable to see intrinsic values as love to be instrumental as emotional assistance for reproductive purposes.

The law is for others, but also to restrict oneself. This does not imply that you in a situation need to love these bounds and always be ready to comply out of free will. Ulysses bound himself to the mast, not to be free to follow his inclinations. Another way of expressing this is that Ulysses followed his second order preferences in a situation when his first order preferences were contrary to his true interests. To implement constraints in such a situation does not strike one as irrational. Rather this example illustrates a situation where prudence is the best policy. The moralists might complain that a truly good person should be able to resist the Sirens by moral will. However, such an experiment might well be the policy of the true fool.

Kant expressed doubts that anybody ever does the right thing just to obey the Moral Law: "I am willing to admit that most of our actions are in accord with duty; but if we look more closely at our thoughts and aspirations, we come everywhere upon the dear self, which is always turning up, and it is this instead of the stern command of duty (which would often require self-denial) which supports our plans. One need not be an enemy of virtue, but only a cool observer who does not confuse even the liveliest aspiration for the good with its actuality, to be sometimes doubtful whether true virtue can really be found anywhere in the world." (Kant 1997, p 164). It is

worth some reflection that the most prominent philosopher for rules of categorical kind is so doubtful about the mere existence of true virtue.

Several researchers have theoretically and empirically showed the desired social effects of trust. A cross country study by Knack and Keefer (1997) estimates that the effect of a difference of one standard deviation in trust corresponds to a difference of half a standard deviation in economic growth rate. Considering such effects it might be argued that it is advisable to promote trust, making it a civic virtue to trust one another. Many people see morals and trust as something in the eye of the beholder; in addition to behave morally, my most important contribution might be to expect others to behave trustworthy and moral.

No doubt there are many people whose social lives are diminished because of lack of trust; the isolated paranoid is an extreme but clear case. But the situation becomes problematic when people with realistic perceptions are asked to take a more optimistic view of their fellow men. Is trust just a social construction or a personal attitude or are there more material factors of importance? Even if the social result would benefit by a general increase of trust, such a change might be a personal loss for the uncritical and a very significant gain for unscrupulous hustlers. In my mind it is only justified to promote a trust that corresponds to the social situation, a trust that brings advantages not only to hustlers or preachers but to the people actually following the advice. From this realistic level further improvements can be made in a virtuous circle of improvements in social climate and increases in trust. But to promote trust in spite of an asocial situation is to fool the naive man. The prime criterion to good advice, must be honesty, not optimism.

One line of argument for unfounded trust is that many social projects are impossible to realize in a situation of mutual suspicion. An example is arms-restriction agreements (e.g. discussed by Harris 1986 and Gauthier 1997). Some possible deals can be properly surveyed so that violation can be punished and damage limited, but if arms-restriction agreements were to be limited to prudence, they would not be as far reaching as if the parties could trust each other. If each party's decision is secret and private, they become independent and just prudence would result that both might choose defect. Gauthier is aware of the risks involved, but he still long for reaching further "In other words, each must find it advantageous to insure that their choice of strategies are interdependent, so that the pact will always be prudent for each to keep. But it may not be possible for them to insure this and to the extent that they cannot, prudence will prevent them from maximizing mutual advantage." (Gauthier 1997, p 259)

If your adversary is a gentleman you might expect him to follow a gentleman's agreement, but if your opponent is ready to kill you, it is hardly outrageously pessimistic to think that he might lie to you. There is something to be said for 'better safe than sorry'. The proper way to maintain trust is to avoid burdening it too much. Two hostile nations or individuals will attribute some value to communication. If this communication is not to dissolve into tiring quarrels, some respect and honesty have to be put into the relationship. Suspicion sets some limits to the amount of trust present in the relationship, and if this limit is properly set no one is offered a 'golden opportunity' of cheating that will give him an advantage that exceeds the loss of goodwill. To expose a semi-friendly relationship for such a golden opportunity might

not only be seen as a naive belief in morality, but as a step to destruction of the relationship. The conclusion of this reasoning is to support trust when it is likely to mean trustworthy, but to be skeptical when trust is just faith.

A paradox is that morals influenced by consideration about compliance more often are criticized for having problems with obtaining this goal, while morals less concerned with compliance succeed in avoiding such critical considerations. There is a widespread opinion that rational morality has more severe problems with compliance - maybe because at heart, morals are seen as an extra constraint, and a rational consideration is no extra constraint but rather foresight or wisdom. It seems to me that proponents of pure virtue disregard the question of compliance. Just the statement that normative principles should be manifested in deeds, not only words, is not enough. Philosophers see themselves more as law makers than law enforcers. And the prospect of letting enforcement consideration influence what is to be considered right is seen as arguing in the wrong direction. However, disregarding a problem is not solving it and not even a good way of avoiding it.

6 Rationality and punishment

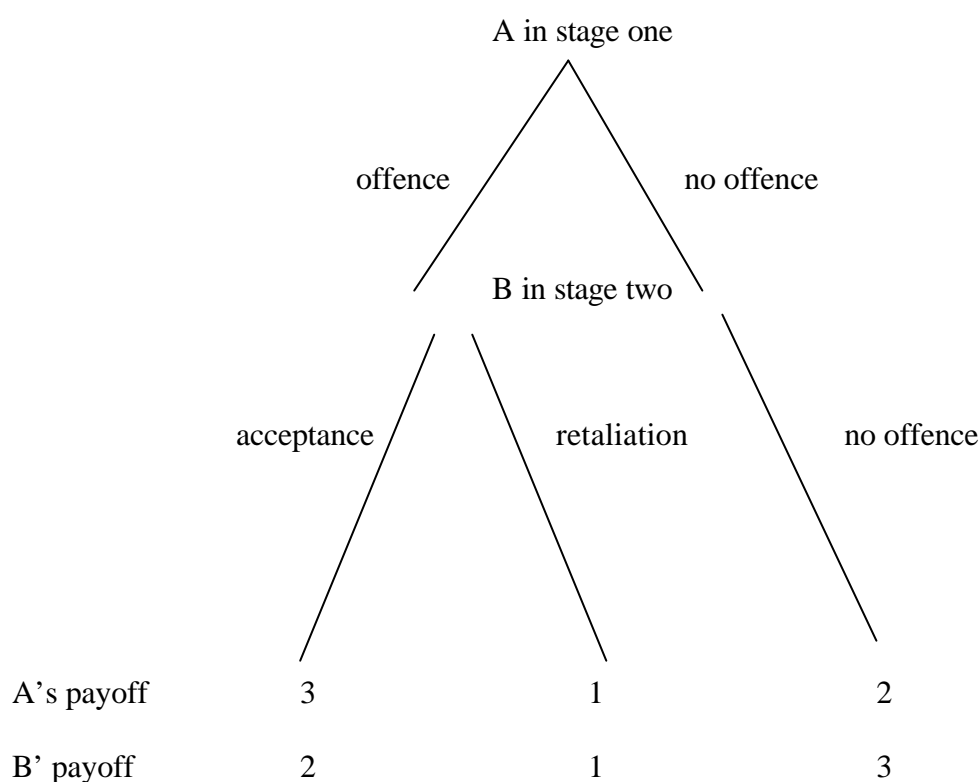
Straight rationality is not only criticized for deficits in bringing the agent herself to comply, but also for providing weak reason for her to push others to behave according to morality (Sen 1987, p 73). Clarifying communication is a proper tool for everyone, but if you are an act egoist your possibilities of influencing others become limited to warnings and assurances. You can solve coordination problems by clarifying your next move, and if this move is in line with your interest, the other person has no good reason to doubt that you will also walk the talk.

In game theory threats and promises are called strategic moves and imply that a person B by promises or threats makes a commitment in stage 0 about potentially doing some kind of sacrifice in stage 2. B's behavior in stage 2 is conditionally linked to what A will do in stage 1. From strict act egoism it is rational to make the promise or threat in stage 0, to influence A in doing something favorable to B, but not rational for B to actually fulfil them in stage 2, regardless of A's behavior in stage 1. However, if the inter actor A is aware of this, he will discard B's promises and threats as 'cheap talk' and expect B to do what is best for him in the actual stage 2 situation. If B does not want to limit his influence he has to, at least occasionally, implement actions that are costly to himself to back up his promises and threats. (Dixit, & Nalebuff 1991)

To influence others the person has better adjust his policy from strict act rationality to looking at the package (stage 0 - 2) and social interaction as series of packages. Many philosophers hold that a more long term strategic consideration is not sufficient to back up the commitment in stage 0 and want B's behavior transformed from calculated strategic move to categorical morality, that is strict deontological behavior independent of long-term consequences. I will discuss two different lines: the firm irrational commitment by Frank (1988) and the firm rational commitment by Gauthier (1986).

The problem description needs to be developed a little more. A has a choice in stage 1 between 'offence' and 'no offence'. In stage 2, B will also choose 'no offence' if that is A's choice. But a choice of 'offence' by A can be followed either by 'acceptance' or by 'retaliation'. The different outcomes can be described as payoffs on an ordinal scale. A's payoff of outcomes (A's action; B's action) is: 3 (offence; acceptance), 2 (no offence; no offence) and 1 (offence; retaliation). B's payoff is: 3 (no offence; no offence), 2 (offence; acceptance) and 1 (offence; retaliation). What can B do to stop A from offence and then expect B to verbally object, but then rationally choose a bad alternative rather than the very worst?

Figure 5



The outcome (offence; acceptance) is what is to be expected according to Reinhard Selten's theory of subgame-perfect equilibrium (1975). The division of the problem into this sequential subgame produces a result that B is unhappy with, but it is hard for him to get out of this current of events. Is there a way out?

Frank argues for irrationality and uses as example a farmer whose land is intruded upon by another farmer's cattle. The farmer tends to accept this if rational, since it is costly to go to court and the damage by the offence is limited. Only if he can feel a strong personal offence or a holy rage will he be ready to make the sacrifices for punishing the offender. Frank does not think it is enough with "Smith's carrot and Darwin's stick" (Frank 1988 p 249). Justice needs irrationality. In many situations the retaliation example is very unattractive, and both rationality and prudence will advise against it. The only effective threat is one built upon a moral indignation that

is not rational when considering its own inconvenience. The desired social solution is obtained if A can read B's moral character and withholds the offence.

There can be several reasons hindering B from striving for justice. He might be scared and scarred or just abstain out of convenience. One important solution is the juridical system that carries most of the burden of keeping A in line with the agreed morality. The citizens can also agree on taking on a personal burden to protect the system as an obligation to comply as a witness. To a high degree the responsibility of retaliation is taken by the system of justice and the cost is shared between citizens. On important issues it is both rational and practically possible to maintain justice by such a limited burden on the citizens, so they will comply. I see no proper place for irrationality.

Gauthier argues that it is not enough to use strategic moves, but instead one should make a moralistic firm commitment. Gauthier disapproves of irrationality as well as of uncertainty and effects on the personal reputation as arguments for the commitment. The behavior in the situation should be motivated by facts in this situation alone and not by further considerations. Gauthier argues for a calculation of a rational morality for these stages taken as a package. Morality orders a categorical implementation of the action justified by reason. Then the case is closed. The consequences of the worst alternative in stage 2 - even if most unfavorable - should not be disbanded.

Gauthier (1997) takes a drastic but most instructive example to make his point: the retaliation to consider is full scale nuclear war. Gauthier sees the rational justification as a calculation of the effect of the threat. The expected value of threat with the possibility of averting offence and the risk of retaliation is compared with the base case of B's medium alternative (this being the preferred alternative of the adversary). Gauthier wants to evaluate the retaliation package and if found rational will go for it. If assured destruction has a positive effect by significantly reducing adversary aggression at a low probability of bringing the world to a nuclear disaster, the retaliation package should be taken as a rational policy. If the adversary then chooses offence despite the dire threat of nuclear retaliation, the threat should be carried through. I think this drastic example deserves praise for being clarifying instead of being dismissed as revolting.

Gauthier's reasoning about retaliation is sound up to a point; rationality gives a stronger foundation for a rule. A position is normally also emotionally weaker if it lacks rational support, as Gauthier writes: "It is because we can give morality a rational basis that we can secure its affective hold" (1986, p 339). But I differ in opinion from Gauthier when it comes to closing the case after the calculation at stage 0. One might say that such an openness gives an option for offence by lowering the expectation of retaliation, but still such a second thought might not be all negative. There are two decision points and to let the second be automatic might be unwise. Naturally, A is interested in a good predictability not to make a miscalculation of B's behavior, and B wants to give his commitment maximum weight. The suggestions of Frank and Gauthier are linked to the idea that a strong moral virtue is a trait of character that is rather transparent so that the virtuous person will not be damaged in never-ending combats, but instead often rewarded by virtue as an effective deterrent against offence. This is reasonable if seen as a disposition and a probability, but

unjustified as certainty. Few things in life are absolutely predictable and when a human being stands in front of a catastrophe, or a golden opportunity, or a situation that is exceptional, she will be ready for a re-evaluation; she will think over her alternatives. This possibility will to some degree always be present in the expectations of others regardless of holy commitments by the agent. There are numerous suggestions for absolute moral laws that offer no exception at all. But, I do not believe there is such a thing as an absolute moral law for the conduct of real people, nor that we can expect other people to believe that we follow such a law. There are always doubts, reasonable doubts.

Of course a strong preventive effect would be attained if all potential murderers or aggressive nations knew with 100% certainty that there would be retaliation. I can see the superiority in theoretical effect when retaliation is linked to a doomsday machine or an infallible God. I still persist that morality has to be sustained by something that might not be best theoretically, but practically good enough. Between the no retaliation and the one cast in stone there is a reasonable alternative that seems likely considering interests, costs and emotions. A more realistic aim should be to create a negative expected value for undesired behavior.

Bentham suggested a formula for prevention of crime with three factors: the punishment of the crime, the detection and conviction rate, and the criminals' knowledge of those effects. (Beckstrom 1993, p 54). I concede that it is likely that criminals and some political leaders are risk takers, so some violations will be carried out even if an action has a negative expected value. Some optimistic bank robbers will indeed test their luck, even under the circumstances that the common and correct understanding is that bank robbery does not pay. This suggested design will not prevent crime, but it will contain it, because most potential bank robbers will abstain or get punished, and the honest man does not feel like a sucker; his own choice does not look irrational because there are a few successful bank robbers. But if this activity is generally profitable it is most demoralizing. The high goal of categorical rules and morality as its own reward is just an ambition - and when failing in rationality social rules are likely to be undermined.

7 Morality as a prudent policy

An adage says: "The road to hell is paved with good intentions". This seems to be especially true when examining morality. The prudent is under siege by the holy, the hopeful and the hypocrite. But justice and morality could instead be seen as less sacrosanct and more practical. Following a rule is not to say that everything else is of secondary importance, but rather believing that the rule will prove its rationality in the long run and with side effects included. The agent accepts that there are exceptions, but also acknowledges that it is hard to find these exceptions, so the practically best policy is to follow the rule. Obeying of rules is supported not so much by principled stubbornness as by the rationale that the potential exceptions are easily misjudged. On many occasions there are pretty clear short-term advantages with lying, but later, when the costs of lying show up, most of us conclude that it is better to resist the convenient temptation to lie. This is not necessarily love for truth, but a conviction of truth as a good policy. Most of us do not pick each other's pockets; we are so convinced that we do not notice, even less take advantage of,

possibilities of pick pocketing that a man of the trade would consider 100 % safe from detection even for an amateur. This view of morality is how a conservative looks upon tradition; some minor advantages do not provide sufficient incentive for really considering breaking the rule.

Many moralists envision a goodness implying that the 'evil' alternative is not even perceived in situations with more significant gains or more dire threats, but this is not realistic. In important situations the special circumstances will be evaluated. If morality is seen as a policy, it seems reasonable to consider alternatives in such situations. However, the practical behavior is hardly different if morality is seen as a virtue instead of a policy. It might sound stringent to pursue a Kantian defense that it is never right to lie, so men should never lie. But it should be questioned whether this really is virtue, and not virtue as a strategy of appearance. Even if a person claims to be an absolute believer, there are no good reasons to believe he is. Just the multitude of attractive norms and honorable virtues leads necessarily to some relativism. Conflicting categorical imperatives, weakness of will, evident absurdities all shake the firm belief. Absolute morality is just an illusion or a lie, but to pledge for it and pursue it to some degree is a policy.

Rational evaluation of all alternatives and their consequences is also an unavailable alternative. What is within human reach is to be an act egoist, who gives low consideration to conventions, restrictions and long-term effects. Such a policy might, as the reciprocal, be evaluated according to what degree it is personally and socially successful. One analytical criterion for such an evaluation is whether life should be seen as a string of independent Prisoner's Dilemmas or rather as series of iterated Prisoner's Dilemmas.

Moral policy consists of practicing some rules without seeing them as weak rules of thumb to be abandoned for slight reason, nor as absolute rules to be implemented even if there are very strong indications of a golden opportunity or a high cost. Rigidity might be an unflattering but apt description of moral rules as policy. Much can be said for such moral rigidity, for a reluctance to change one's rules of conduct; all situations are special in some respects but few are really exceptional. I do not think this is just an alternative; that is what moral rules are. Some claim they are absolutes and others claim a permanent pragmatic openness. The real choice is between different kinds of rigid policy. I see three major alternatives: 1 following conventions that are in line with self-interest; alias rule-egoism, constrained maximization or reciprocity. 2 Being less rule oriented and act according to act-egoism; alias 'straightforward maximization' and 'rational' in economics and game theory. 3 Promoting and sometimes following moralistic rules claiming disdain rather than regard for one's own self-interest; as altruism, Kantianism and utilitarianism.

If a specific norm stays in the square of pure morality, that implies that the norm is half-hearted or immature. I would rather see morals following a three-step sequence: moral discussion, moral agreement and prudence. To convert a moral suggestion first into a social agreement and then to prudence should be the goal of the moral process. If an agreement is not reached, or the people are not ready to support the agreed moral with incentives, the suggestion is simply not considered good enough. There are good reasons to be choosy. Too many rules or too heavy-handed support might

undermine the support and shift the rules to pressure, and this is certainly worse than having the rules as pure morality. All societies have some unpopular laws, but there is a limit to what any society can take. It is sometimes said about authoritarian regimes that they are inclined to be content by executing pressure; the people do not have to agree, simply to obey. In contrast, totalitarian regimes have more far reaching ambitions and need more compliance, including pure morality. The soldier Lei Fung, the worker Stachanov and Hitler Jugend Quex were all - according to mythology - martyrs that gave everything to the cause without asking for selfish reward.

For other political philosophies there is less need for self-sacrifice. But for all systems, there is a need to move from pressure to prudence. To get this moral support, power has to be adjusted to moral agreement. However, I am inclined to see morality as an intermediate step to legality rather than strong enough as an isolated factor. Hume had a similar view. He pointed at the strong attraction of receiving obedience from other people and a strong desire to keep that power. A factor that increases the leaders' possibilities to stay in power is whether the rules are seen as just. The leader then has a strong self-serving reason for justice and the ordinary people will have the brute force of authority as a strong reason to obey justice (Hume 1777).

To some people, statements such as "it should pay to be moral" sound paradoxical or at any rate dubious. I have tried to give some reason why one should be suspicious about the opposite view, namely that morals should be a burden; proud and self-sacrificing. Such goals might be more suitable for showing off an attitude than as guidelines for real behavior. If you advise people on real behavior, it is a major advantage if the advice is good in a material sense. If it is not, it also undermines the advice in a moral sense. It seems that prudent morality is less of a dubious goal.

Acknowledgements

I want to thank Hans De Geer, Germund Hesslow and Ingolf Ståhl for valuable comments and suggestions.

References

- Beckstrom, J. H. 1993. *Darwinism Applied. Evolutionary Paths to Social Goals*. Praeger, Westport CT.
- Binmore, Ken. 1994. *Playing Fair - Game Theory and the Social Contract*. The MIT Press, Cambridge, Massachusetts.
The MIT Press, Cambridge, Massachusetts.
- Binmore, Ken. 1998. *Just Playing - Game Theory and the Social Contract*. The MIT Press, Cambridge, Massachusetts.
- Browning, Christopher. 1992. *Ordinary Men - Reserve police battalion 101 and the final solution in Poland*. HarperCollins, New York.
- Dawes, Robyn, McTavish, J. & Shaklee, H. 1977. 'Behavior, communication, assumptions about other people's behavior in a common dilemma situation' *Journal of Personality and Psychology*, 35, 1.
- Dixit, Avinash & Nalebuff, Barry. 1991. *Thinking Strategically - the competitive edge in business, politics and everyday life*. W.W. Norton & Company, New York.
- Drucker, Peter. 1939/1995. *The End of Economic Man - the Origins of Totalitarianism*. Transaction Publishers, London.
- Frank, Robert H. 1988. *Passions within Reason*. W.W. Northon & Co, New York.
- Frohlich, Norman & Oppenheimer, Joe 1992. *Choosing Justice - An Experimental Approach to Ethical Theory*. California University Press, Berkeley.
- Gauthier, David. 1986. *Morals by Agreement*. Clarendon Press, Oxford.
- Gauthier, David. 1990. *Moral Dealing - Contract, Ethics and Reason*. Cornell University Press, London.
- Gauthier, David. 1997. in Rogers, Kelly (ed.) *Self-Interest*. Routledge, New York.
- Harris, C.E, 1986. *Applying Moral Theories*. Wadsworth Publishing Company, Belmont, California.
- Hobbes, Thomas. 1651/1981. *Leviathan*. Penguin Books, London.
- Hume, David. 1777/1992. *An Inquiry Concerning the Principles of Morals*. Oxford University Press, Oxford.
- Kant. 1997. Foundations of the Metaphysics of Morals Section II §407. In Rogers, Kelly (ed.) *Self-Interest*. Routledge, New York.
- Knack, Stephen & Keefer, Philip. 1997. "Does social capital have an economic payoff?" in *Quarterly Journal of Economics* no 4.
- MacIntyre, Alastair. 1981. *After Virtue*. University of Notre Dame Press, Indiana.
- Ruse, Michael. 1986. *Taking Darwin Seriously - A Naturalistic Approach to Philosophy*. Blackwell, Oxford.
- Selten, Reinhard. 1975. 'Reexamination of the perfectness concept for equilibrium points in extensive games.' in *International Journal of Game Theory* 4:25-55.
- Sen, Amartya. 1987. *On Ethics and Economics*. Basil Blackwell Ltd, Oxford.
- Österberg, Jan. 1988. *Self and Others*. Kluwer Academic Publishers, Dordrecht.