

# Moral Compliance and the Concealed Charm of Prudence

*Jan Tullberg*

**ABSTRACT.** The key to moral behavior is often perceived to consist of ignoring rational self-interest and instead following norms recommended by religious tradition and moral philosophy. A central issue is the connection between these ambitions and actual behavior. Are an idealistic mood and an ethics of ambition the way out of an iron cage of individualistic rational behavior? Or is ethics best served by rules and incitements in harmony with rationality? The article discusses morality from the perspective of compliance. A normative suggestion in the Prisoner's Dilemma exemplifies the case of prudent morality. The player should contribute if the expected value of that choice exceeds his payoff of mutual defection. 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 favoring a good average result in the long run is the most relevant argument for complying with a rule. The structure of economic games can give useful insights about problematic situations and consequences of different strategies. Being prudent rather than doing good might be a better policy not only for the agent, but also for organizations and society.

**KEY WORDS:** compliance, incentive compatible, moral policy, prudence, rationality

**JEL CLASSIFICATION:** A13

## 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 philosophical thought.

Hobbes (1651) termed the central moral agreement the "Second Law of Nature," but the main theme of the article is to penetrate the problem Hobbes called the "Third Law of Nature." That is, it is not enough to reach a moral agreement – there is also a need for compliance with this moral agreement.

Since there is some interdependence between morality and compliance, different moral principles are most relevant for the discussion. In this article, the chief candidate is the proposition 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 mold such as 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 such a choice and investigate the case for prudence.

Many issues of rules and morals at the macro level of society have similarities with issues at the meso level, the company and its organizing of employees and interacting with outside stakeholders. Visions of the good company have similarities with visions of the good society. Many persons think the companies should be more caring, while other persons worry about a widening gap between an ethics of rhetoric and an ethics in use and also perceive intellectual limitations in the suggested new nice ways. Are altruism, deliberative democracy, an ethics of care and environmental stewardship the elements of a better moral framework? Organizations like EABIS

(2007) strongly promote the thesis that such ideas will be good for companies, for societies, and for career interested students. They will also be beneficial for the business schools that educate the new virtues; appearing to imply a win–win situation for all. Educational literature offers support for such an approach (Carroll and Buchholtz, 2003; De George, 1999; Fisher and Lovell, 2006). In this article, the attitude is less gung-ho and focuses on another kind of sustainability, ethical sustainability. Hobbes questions are central and the article will contribute with an answer considering skepticism to idealistic ethics.

### Different kinds of compliance

A classic discussion focuses on compliance, prudence, and morality. 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 morality 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 – incentive support as reward or punishment. It might be instructive to see four combinations of these two major components (Figure 1).

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

Figure 1. Different motives for compliance with a rule.

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 explained differently from prudence, since the agent acting on the moral idea often could obtain 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 negative or positive incentives outweigh the agent’s own moral conviction, he complies with the social incentives.

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 either the one that 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 customs that have lost their *raison d’être*.

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

It might be argued that the category of pressure is the antithesis to pure morality, but it is usually the dubious supporters, not the opponents, that are scrutinized and separated from pure morality. The “dubious supporters” include people who comply only when observed and those who praise with words but deviate in deeds. 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 being 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 (e.g., Kant, 1997).

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 rules with more support, but also fewer rules. Prudent morality is a morality of performance rather than a morality of ambition.

Generally a prudent morality will be focused and advocate negative rules. Rules are primarily important restrictions in a more general quest for what the agent considers to be good. Hippocrates advocated doctors’ obligations toward patients, but the central rule he proposed was “Do no harm.” This is in line with the Ten Commandments containing eight “thou shalt not” principles and just two “thou shalt.” Moral rules are primarily constraints, not positive constitutive rules.

### **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. The dilemma is built on a narrative. You and your partner in crime have been caught by the police and are now interrogated separately. If you confess and blame your partner, you will be set free. If you say nothing, but your partner describes you as the prime criminal you will end up with a long stay in prison. If none of you give the police any information, the evidence will only give each of you a short stay in prison. If both of you talk, there is more evidence resulting in a medium prison stay. Your partner has the same choice in another cell and your choices will be made simultaneously. Should you confess to the police or not?

As mentioned, I will adopt a position in the normative question of how one should behave in such a Prisoner’s Dilemma situation. The reasoning will be illustrated with calculations from the following matrix.

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 in the matrix, not for an action. Instead the positive action is called “Contribute.” This might result in cooperation, but not necessarily so; it takes two to tango.

The game-theoretical judgment is that a rational person should always choose to defect since this gives the best result regardless of the other person’s choice of action. If A (referred to as male) chooses defect, he will get 4 compared to 3 if he plays contribute, assuming B (referred to as female) plays contribute. If B plays defect, A will get 2 instead of  $-1$  by choosing defect. Therefore, the players will end up in the non-cooperation outcome of (2; 2). Non-cooperation becomes a Nash equilibrium implying that none of the players will change strategy. If one player deviates from defect and chooses contribute instead, he/she will get impairment to  $-1$ . This result 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 if 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 variants where 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 the example of Figure 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 she will

		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)

Figure 2. Prisoner's Dilemma.

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 only 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,” which emphasizes other aspects. This is a prudent policy combining incentives (a good but not an optimal payoff) and morals (of mutual obligations not of unselfishness). The ambition is to maximize long-term sustainable cooperation.

This conditioning of one's behavior upon the expected behavior of the other agent might bring objections from advocates of a more categorical or universal morality, but from a contractarian point of view it sounds reasonable. If morals are mutual agreements, it is natural that behavior is influenced by tacit or explicit rules and, more ultimately, actual behavior of 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 long-term rationality, not as a force countervailing reason and interest. The Darwinian effects of kin selection, which modify individual interest, are aspects not incorporated in this article.

Frank (1988) presents an experiment with the Prisoner's Dilemma in which the test persons were asked about their expectations of the other person's behavior. Of those who thought the other person would contribute, 83% decided to contribute. Among persons who thought that the other person would defect, the share of contribute was only 15%. Work by Dawes et al. (1977) shows the same pattern. This experiment also finds that one explanation for the difference is that the expectations partly are rationalizations of the chosen decision. Such behaviors indicate normative support for a symmetric outcome. People do not want to be either suckers/saints or free-riders – they want to reach cooperation if there is a good possibility for success, and non-cooperation when the other person is not contributing. 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 to 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 variants – this is normally classified as “cheap talk” (Farrell, 1987). 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 reputations, he might get a significant improvement by fooling some of the people some of the time. Fifteen percent 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 a manner similar to the reciprocal agent burdening the contribute alternative with the risks of

defection from the co-player, the Foole sweetens the defect alternative with own free-riding gains. 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 judgments will distinguish between a person abstaining from cooperation and a person defecting from given promises. If declaring no intention for a cooperative outcome, 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 Prisoner's Dilemmas, it is not a problem to claim that the reciprocator can do better than the person starting with a gain by defection. As a sequence of four outcomes, the cooperative "second best, second best, second best, second best" beats a straightforward "best, third best, third best, third best" (if using the payoffs of this example, 12 points vs. 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, 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 can only try to be one. You can decide between the strategies of robbing banks by hold-ups and by nightly break-ins, but you still cannot 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 no 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 who 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 about opportunists following strong extrinsic incentives of moral ideas which few sympathize with today, those of 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 (Browning, 1992). How much of a divergence is caused by a lack of intrinsic belief? The outcomes are not directly due to the intrinsic belief of the agent, but to his choice of defect or contribute.

The personal outcome for the individuals of different strategies might be rather similar. The act-egoist will sometimes obtain extra results by a cunning defection and sometimes lose by causing a strong negative reaction. The mixture of altruists, reciprocators, and opportunists in society indicates that all strategies have possibilities. The thesis suggested here is that a reciprocal society is a better society because of a higher level of cooperation. If so, it is of interest to consider possibilities to promote cooperative behavior.

### **Sustaining reciprocal morality in the Prisoner's Dilemma**

If our reciprocal person is unaffected by the temptation of the Foole, there are developments of the present situation that might influence a change in behavior. One threat is that many other reciprocators 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 reciprocators in the population, it might be motivated to discriminate against all persons who are not confirmed reciprocators.

A contrary development is to introduce a policy to reduce the profitability of free-riding. If reciprocators 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 payoff reduced 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 contribute and our reciprocators will get a better situation, to such an extent that it will favor 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 on 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 the risk of destructive rules should be addressed. Some people will advocate upgrading of some ideas – old traditions or new trends – presently dwelling as Soft Conformism. Others would propose bringing some honored ambition of pure morality into prudence by making it not only desired, but also required. A new restriction might be considered social progress in the judgment of some citizens, but perceived as steps in the wrong direction by others. The heated discussion in many issues indicates that there are seldom evident, non-controversial, solutions. Still, there is no reason indicating a more serious problem such as a trend to decay – a tendency to choose misguided rules and incentives more often than changes successfully pushing behavior in the right direction. Rather, trial and error would direct efforts curbing

problematic asocial behavior towards more important goals and more efficient methods. Now let us turn our attention to other suggestions for obtaining the cooperative solution in the Prisoner’s Dilemma. Instead of an improved support structure, “higher” moral norms that overrule self-interested concerns are seen as the crucial factor.

### 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 implicate 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 she thinks 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 sinners triumph over the true faith (Tolerance; Holy War). Payoff  $-1$ .

According to Person B, Jehovah’s ranking order, the outcome (Holy War; Tolerance) and the outcome (Tolerance; Holy War) trade places, and the *déjà vu* suspicion is confirmed (Figure 3).

As documented in history, it is also possible to believe in the same God and the same Holy Scripture, but still to 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

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

Figure 3. Prisoner’s Dilemma: altruistic conflict.

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, and not only in religious circles, is to turn the other cheek and suffer, rather than taking an eye for an eye. This approach suggests less focus on 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, a thesis that 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 play contribute even if the other player does not.

For an analyst of games, this sounds like very dubious advice. Apart from being costly to the agent deliberately acting saintly, there is a further problem. When one agent becomes a saint, the other agent can become a free-rider. Regardless of the generous agent’s intention, she is then supporting asocial behavior. Even if ignoring her own cost, she will, by accepting and supporting free-riding, undermine the expected value for other agents of playing contribute in future games with this pampered player. Later this might influence responsible individuals who are reciprocal, not altruistic, to reconsider their strategy to play contribute. Saintly behavior is often described as a step from non-cooperation toward cooperation, but seen in its social context it is often more likely to be a step from cooperation to sucker/saint and then to non-cooperation. Furthermore, is there not more to say in favor of the symmetric

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

Figure 4. Prisoner’s Dilemma: breaking away.

non-cooperation than for asymmetric transfer from an altruistic giver to an egoistic receiver?

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 your own good, you produce more social good than if you make efforts to help your fellow man. The payoffs of the upper left and the lower right square might trade places (Figure 4).

Greedy consumers will, in their effort to reap value for themselves, put hard pressure upon producers for lower prices and improved products. Fulfilling these demands and still earning a profit is likely to bring more benefits to society than selfless contributions to charity. In the short term, companies might protect themselves successfully against consumer demands by cooperation with cartels stipulating prices or partition of the market through negotiations between companies. But in the long run the better payoff might be found in playing defect and striving for survival in a competitive equilibrium rather than collaborate with colleagues at moral and legal risk.

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 other people’s preferences. With assumptions or higher authorities stipulating preferences, there is less influence for an amalgam of genuine 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.

### The attraction of pure morality

One line of critique is that an attraction to prudence is 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 the choice from an ethical decision to a pragmatic maximizing decision. After such a change in motivation, there will be a weaker defense hindering defection if the incentives or the controls over incentives are insufficient. Then an instrumental treatment of morality undercuts the function of morality as an effective constraint upon behavior (Frohlich and Oppenheimer, 1992). Since long ago, there have been worries about a "crowding out" effect of higher motives by lower motives. For an interesting contribution in this debate see Frey (1994). Perhaps there is a correspondence with Gresham's law, which stipulates that coins with low silver/gold content will push coins with a higher content of high-value metals out of circulation. Like 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 assigned 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 perspective it is reasonable to see intrinsic values such as love to be instrumental as emotional assistance for reproductive purposes.

The law exists primarily to protect from misdeeds performed by others, but sometimes also to restrict the self. This does not imply that in a given situation you need to love these bounds and always be ready

to comply out of free will. Ulysses bound himself to the mast so as not to be free to follow his inclinations. Another way of expressing it is that Ulysses followed his second-order preferences in a situation where 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 shown the desired social effects of trust. A study by Knack and Keefer (1997) compared levels of trust and economic growth in different countries. They concluded that the effect 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 behaving morally, my most important contribution might be to expect others to behave in trustworthy and moral ways.

No doubt there are many people whose social lives are diminished because of lack of trust; the isolated paranoid is an extreme but palpable 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 and 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 to hustlers or preachers, but to the people actually following the advice. From this realistic level, further advances 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 of good advice must be honesty, not optimism. If very few checks will bounce, there are only disadvantages with being suspicious. On the other hand, if bad checks become frequent, there is no virtue in continuing to be optimistic about your fellow man because that will primarily help the cheaters. If payments with plastic cards are more trustworthy, it makes sense – private and for society – shifting to that medium.

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, the two become independent and prudence alone would have the result that both parties might choose to defect. Gauthier is aware of the risks involved, but he still longs 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 over-ambition. Two hostile nations or individuals will attribute some value to communication and, 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 an advantage exceeding the loss of goodwill. To expose a semi-friendly relationship for such a golden opportunity might be seen not only as a naive belief in morality, but as impairing the relationship. The conclusion of this reasoning is to support trust when it is likely to imply trustworthiness, but to be skeptical when trust is just faith.

A paradox is that morals influenced by consideration about compliance are more often 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 – perhaps 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. Philosophers see themselves more as lawmakers than as law-enforcers, and letting considerations of enforcement influence what is to be considered right is seen by them as arguing in the wrong direction, starting from implementation. Statements that normative principles should be manifested in deeds, not only in words, are still just words and not really addressing the implementation issue. Therefore, I argue that the practical possibilities to maintain a rule should be considered an important factor when analyzing whether a suggested rule is qualified.

### **Rationality and punishment**

Straight rationality is criticized not only 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 before person A makes his move in stage 1. B's behavior in stage 2 is conditionally linked to what A has done. This commitment to a reward or a punishment implies potentially making some kind of sacrifice when B herself will act. With strict act-egoism it is rational to make the promise or threat, to influence A in doing something favorable to B, but not rational for B to actually fulfill the promise or threat in stage 2. However, if 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 her in the actual stage 2 situation. If B does not want to limit her influence she must, at least occasionally, implement later actions that are costly in order to back up her promises and threats (Dixit and Nalebuff, 1991).

To influence others, a person had better adjust his policy from strict act rationality to see the commitment and the two stages as a package. Many philosophers hold that a more long-term strategic consideration is not sufficient to back up the commitment and want B's behavior transformed from calculated strategic move to categorical morality, i.e., strict deontological behavior independent of 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 can be developed a little more into a figure. 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 and actions (A's action; B's action). A's payoffs of different outcomes are 3 (offence; acceptance), 2 (no offence; no offence), and 1 (offence; retaliation). B's payoffs are 3 (no offence; no offence), 2 (offence; acceptance), and 1 (offence; retaliation) (Figure 5). What can B do to stop A from offending and then expecting B to verbally object, but then rationally

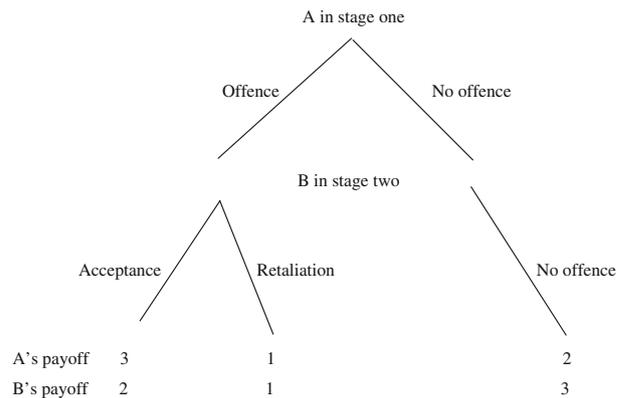


Figure 5. Threats and two decisions in sequence.

choose a bad alternative (offence; acceptance) rather than the very worst alternative (offence; retaliation)?

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

Frank argues for irrationality, and illustrates with 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 to have "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 therefore one built upon a moral indignation that does not rationally consider its own inconvenience. The desired social behavior is obtained if B's moral character is communicated and A therefore abstains from the offence.

There can be several reasons hindering B from striving for justice. She 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. To a high degree the responsibility and the cost of retaliation are shared between citizens. Not all rules are supported by the legal system, but many

minor violations provoke social criticism by “significant others.” In many situations the contract between two agents is incomplete in juridical terms, so more general ideas of fairness become relevant for finding a solution without a conflict. A reputation for fairness is therefore important, and significant others can with little inconvenience influence this reputation. Their perception of an action by an agent as fair or unfair is also a factor determining when these others will choose between interacting and avoiding this agent. On important issues it is both rational and practically possible to maintain justice with such a limited shared burden on the citizens, so they collectively can maintain a pressure for complying. Therefore, I disagree with Frank and see no proper place for irrationality.

Gauthier argues that it is not enough to use strategic moves; instead, one should make a firm moralistic 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 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 should not be excluded.

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 of retaliation. 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 (offence; acceptance), this being the preferred alternative of the adversary. Gauthier wants to evaluate the retaliation threat package and, if finding it rational, decide to go for it. If the threat of destruction has a positive effect, by significantly reducing adversary aggression at a low probability of bringing the world to a nuclear disaster, the retaliation threat package should be accepted as the 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. Indicating such openness increases risks of offence by lowering the expectation of retaliation, but such a second thought might still not be entirely negative. There are two decision points, and to let the second be automatic might be unwise. Naturally, A is interested in good predictability so as not to make a miscalculation of B’s behavior, and B wants to give her commitment maximum weight. The suggestions of Frank and Gauthier are linked to the idea that a strong belief in retaliation as a virtue is a trait of character that is rather transparent, so that the virtuous person will not be damaged in frequent combats, but instead often rewarded by her virtue functioning 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, or 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 insist that morality has to be sustained by something that might not be theoretically best, but practically good enough. Between no retaliation and the one cast in stone, there is a reasonable alternative that seems most plausible in view of 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 some 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. The suggested design with negative expected value 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 robbing banks is a generally profitable activity, this is most demoralizing. The high goal of categorical rules and morality as its own reward is just a hypothesis. Social rules are likely to be undermined when failing in rationality.

In a series of public-goods experiments, Fehr and Gächter (2000, 2002) found that participants did not contribute fully regardless of whether others were cheating or not. Instead, participants in the iterated game decreased their contributions as others decreased theirs, in a vicious circle to the detriment of all participants. In another treatment of the game, with possibilities to punish co-players, the behavior developed very differently. The level of contribution increased to a very high level. No less than 84 percent of participants punished another player at least once during the experiment, despite this being at a cost to themselves. It is reasonable to see this as an inclination to punish asocial behavior, to bring people up to normal social standard. Free-riding and parasitism are serious offences that can be countered efficiently. In contrast, supererogatory acts are costly for the agent and their social influence and benefit are often dubious.

Many thinkers perceive a conflict between rationality and the burden of punishment. The reasoning around these examples indicates that this is not a crucial problem.

### **Principles and 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 grants 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 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 pickpocketing which a man of the trade would consider safe from detection even if performed by a novice. This view of morality is how a conservative looks upon tradition; some minor advantages do not provide sufficient incentive for really considering change.

Many moralists envision a goodness implying that the "evil" alternative is not even perceived in situations with significant gains or 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 extreme 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 in practice, and not just virtue as a strategy of appearance. Even if a person claims to be an absolute believer, there are no good reasons to believe she is. The very 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 the degree in which 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 the slightest reason, or as absolute rules to be implemented even if there are very strong indications of a golden opportunity or staggering costs. 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, but what moral rules substantially are. Some claim they are absolutes and others claim a permanent pragmatic openness, but the real choice is between different kinds of rigid policy.

### **Choice of moral 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 behaving 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; according to altruism, Kantianism, and utilitarianism.

If a specific norm stays in the square of pure morality, this implies that the norm is half-hearted or immature. I would rather see morals following a three-step sequence: moral discussion, moral agreement, and prudent morality. 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 if people are not ready to support the agreed moral with incentives, the suggestion is simply not considered good enough. There are valid reasons to be choosy. Too many rules or too

heavy-handed incentives 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 demand moral agreement, including pure morality. The soldier Lei Fung, the worker Stakhanov and Hitler Jugend Quex were all – according to mythology – martyrs who 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 pure morality as an intermediate step rather than strong enough as motivation. David Hume had a similar view. He pointed at the leadership's interest in obedience by other people and in a strong desire to keep power. A factor that increases the leaders' possibilities to stay in power is whether the rules are seen as just. The leaders then have 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 self-sacrificing burden. Such morals 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. According to this analysis, prudent morality stands out as more solid and sustainable, even though its charm is not sparkling but concealed.

### **Acknowledgments**

I want to thank Hans De Geer, Germund Hesslow, and Ingolf Ståhl for valuable comments and suggestions and Jon van Leuven for improving the text.

## References

- Beckstrom, J. H.: 1993, *Darwinism Applied – Evolutionary Paths to Social Goals* (Praeger, Westport, CT).
- Binmore, K.: 1998, *Just Playing – Game Theory and the Social Contract* (The MIT Press, Cambridge, MA).
- Browning, C.: 1992, *Ordinary Men – Reserve Police Battalion 101 and the Final Solution in Poland* (HarperCollins, New York).
- Carroll, A. and A. Buchholtz: 2003, *Business and Society – Ethics and Shareholder Management* (South-Western Thomson, Mason, OH).
- Dawes, R., J. McTavish and H. Shaklee: 1977, 'Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation', *Journal of Personality and Psychology* **35**(1), 1–11. doi:10.1037/0022-3514.35.1.1.
- De George, R. T.: 1999, *Business Ethics*, 5th Edition (Prentice Hall, Upper Saddle River, NJ).
- Dixit, A. and B. Nalebuff: 1991, *Thinking Strategically – the Competitive Edge in Business, Politics and Everyday Life* (W.W. Norton & Company, New York).
- Drucker, P.: 1939/1995, *The End of Economic Man – the Origins of Totalitarianism* (Transaction Publishers, London).
- EABIS: 2007, '2nd Special Report on Corporate Responsibility and Global Executive Education', <http://www.eabis.org/resources/2007ExecutiveEducationSpecialReport>.
- Farrell, J.: 1987, 'Cheap Talk, Coordination, and Entry', *The Rand Journal of Economics* **18**(Spring), 34–39. doi:10.2307/2555533.
- Fehr, E. and S. Gächter: 2000, 'Cooperation and Punishment in Public Good Experiments', *The American Economic Review* **90**(4), 980–994.
- Fehr, E. and S. Gächter: 2002, 'Altruistic Punishment in Humans', *Nature* **415**, 137–140. doi:10.1038/415137a.
- Fisher, C. and A. Lovell: 2006, *Business Ethics and Values – Individual, Corporate and International Perspectives* (Pearson Education, London).
- Frank, R. H.: 1988, *Passions Within Reason* (W.W. Norton & Co, New York).
- Frey, B. S.: 1994, 'How Intrinsic Motivation is Crowded Out and In', *Rationality and Society* **6**(3), 334–352. doi:10.1177/1043463194006003004.
- Frohlich, N. and J. Oppenheimer: 1992, *Choosing Justice – an Experimental Approach to Ethical Theory* (California University Press, Berkeley).
- Gauthier, D.: 1986, *Morals by Agreement* (Clarendon Press, Oxford).
- Gauthier, D.: 1990, *Moral Dealing – Contract, Ethics and Reason* (Cornell University Press, London).
- Gauthier, D.: 1997, 'David Gauthier', in K. Rogers (ed.), *Self-Interest* (Routledge, New York), pp. 253–265.
- Harris, C. E.: 1986, *Applying Moral Theories* (Wadsworth Publishing Company, Belmont, CA).
- Hobbes, T.: 1651/1981, *Leviathan* (Penguin Books, London).
- Hume, D.: 1777/1992, *An Inquiry Concerning the Principles of Morals* (Oxford University Press, Oxford).
- Kant, I.: 1997, 'Foundations of the Metaphysics of Morals (Section II §407)', in K. Rogers (ed.), *Self-Interest* (New York, Routledge).
- Knack, S. and P. Keefer: 1997, 'Does Social Capital have an Economic Payoff?', *The Quarterly Journal of Economics* **112**(4), 1251–1288. doi:10.1162/003355300555475.
- MacIntyre, A.: 1981, *After Virtue* (University of Notre Dame Press, IN).
- Österberg, J.: 1988, *Self and Others* (Kluwer Academic Publishers, Dordrecht).
- Ruse, M.: 1986, *Taking Darwin Seriously – a Naturalistic Approach to Philosophy* (Blackwell, Oxford).
- Selten, R.: 1975, 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* **4**, 25–55. doi:10.1007/BF01766400.
- Sen, A.: 1987, *On Ethics and Economics* (Basil Blackwell Ltd, Oxford).

Stockholm Centre for Organizational Research,  
Stockholm, Sweden  
E-mail: jan@tullberg.org