

Conventions and Cooperative Virtues

Some Observations on Evolutionary Game-Theoretic Explanations of Social
Norms

Bruno Verbeek

Social & Political Theory, Australian National University

Conventions and Cooperative Virtues

Some Observations on Evolutionary Game-Theoretic Explanations of Social Norms

§1 Introduction

The ubiquity of norms is overwhelming. There are (detailed) norms regulating our behavior in community at large, there are norms that regulate our actions in the school we attend, in the organizations we join, in the workplace we frequent. There are norms that tell us what to wear, how to eat and how much real fruit there should be in orange juice. The sequence of characters on this page is dictated by a norm. The important occasions in our lives ranging from birth to burial are structured by norms. In addition, there are norms regulating property, economic transactions, taxes, and there are norms which form the basic structure of society.

No wonder then that philosophers and social scientists alike always have been interested in norms. The challenge is to provide a theory that explains the nature of our norms. One such theory is *conventionalism*. Conventionalism makes the following claims. First, that it is instrumentally rational to comply with norms. This is *reductionism*: norms are a subset of the set of rational prescriptions. Conventionalist theories are not alone in this adherence to reductionism. For example, Hobbesian contractarianism holds that one should comply with moral prescriptions in so far as they are the result of a hypothetical contract between rational agents. What distinguishes conventionalist theories from other reductionist theories of norms is the second claim of conventionalism: (part of) the reason that this is rational to comply with a norm is because it is known that all, or a sufficiently great number of others in the group, comply with those norms. This is *conformity*.

With the introduction of game theory in the social sciences, conventionalism has been reformulated in game theoretic terms. Though the pioneering work was done in the sixties by Thomas Schelling and David Lewis, the real “boom” is of the last twenty years or so of this century.¹ We have witnessed a plethora of publications that formulate models to explain the emergence and stability of norms. In this paper I will try to determine how

¹ Schelling, Thomas (1960), *The Strategy of Conflict*, Cambridge: Harvard University Press.; Lewis, David (1969), *Convention: A Philosophical Study*, Cambridge: Harvard University Press..

successful the conventionalist, game theoretic models are as characterizations of the nature of norms. In particular, I will raise the question whether the models explain the characteristics of what we intuitively would label as a “norm” or whether something is lacking.

Section 2 presents an informal definition of norms that is derived from the work of Hart (1961; Winch (1990). Section 3 and 4 discuss a simple model of a property norm based on the “hawk-dove” game. Section 5 assesses whether this model can account for possibility of deviance from a norm. Section 6 discusses the role of sanctions in the model. Section 7 argues that sanctions presuppose the existence of moral motives. Section 8 deals with how deviance provokes resentment. Section 9 provides an alternative analysis of resentment, which shows that resentment presupposes moral motives. Section 10 draws some conclusions about the success of game theory in the analysis of social norms.

§2 Characteristics of norms

First, we need to clarify the intuitive concept of norms. In what follows I give a list of features of norms. I am reluctant to claim that it is an exhaustive list. Nor am I claiming that any of these features are necessary or sufficient.² However, if we find the sort of things that have (most of) these features, it is likely that we are dealing with a norm.

To determine whether a group follows a norm, the following phenomena are relevant:

- (1) *There exists regularity in behavior.* This immediately leads to a puzzle. What if nobody behaves as the norm prescribes, yet everyone judges behavior in terms of that norm? A plausible reaction to this puzzle is to claim that in such situations, a judgment in terms of a norm that is generally deviated

² My reluctance is motivated by two considerations. First, the aim here is to characterize norms and then see whether a particular theory or model captures this characterization. Such a theory would formulate a definition of norms. Providing a definition of a norm – as opposed to a characterization – presupposes a theory of norms. However, this is not so terribly important. I am willing to accept that what we intuitively label “norms” is theory-laden in this sense. The aim then becomes to formulate a theory that expresses these intuitions optimally.

My main consideration is that I am not altogether convinced that these features are sufficient to separate what we would intuitively characterize as a norm. A good example is the “rule” in baseball that the third-base man should draw in when he expects a bunt. It satisfies all the criteria but is it what we label a “norm”? Personally, I would be inclined to say it is, but others do not. For example, [Shapiro, 1998 #840].

- from cannot be taken seriously. Such a judgment is more like an unreasonable ritual. It does not follow from an existing norm Bartsch (1983).
- (2) *This regularity can be acquired through learning.* Not all regularities in behavior are norm-guided. Babies sleep a lot, people are repelled by the smell of rotten eggs, and lions defend their young. None of these just mentioned phenomena are norm-guided. They are determined by biological, physical, or other facts of the species. What we are after concerns regularities which are not imposed on us in this manner.
 - (3) *Deviation from the norm is possible.* This feature is connected to the second criterion. Normally, one learns a rule by trial and error, by both conforming to it and by occasionally deviating from it.
 - (4) *Those concerned are able to detect deviations from the norm.* That is, when confronted with deviant behavior, the norm followers are able to recognize it as deviance. Of course this does not have to be perfect. Nor does it have to be the case that those concerned can give an exact formulation of the differences between compliant and deviant behavior.
 - (5) *Deviations are met with reproach.* This reproach can take the form of punishment, as in criminal law, or other (informal) forms of social control.
 - (6) *This reproach can be a reason to change one's behavior.*
 - (7) *Reproach of deviations is regarded as correct or legitimate.* This will result in deviants either exercising self-criticism or denying the correctness of the negative response. Conversely, those exercising criticism will regard their reproach as correct. As we shall see, it is precisely this feature that conventionalist theories have difficulties in explaining.

Summarizing the features of the list, we get the following informal “definition”: when we think of norms we think of regularities in behavior that are socially transferred, and deviations from which can be recognized and will be criticized in some manner, where those concerned will regard this criticism as appropriate.

§3 A model for conventions of property

How does a conventionalist analysis of norms capture these characteristics? Evolutionary game theory has been employed in demonstrating how these regularities emerge and can be stable. In order to appreciate their success as an

explanation of norms, I will present one relatively simple designed to capture (part of) property norms.

Property can be understood as the allocation of scarce resources. The fact that resources are scarce generates conflict between agents who have a use for the disputed resource. Imagine the following situation: two contestants are in conflict over a resource. This conflict is not an all or nothing affair; it is not a zero-sum game. For one thing, one of the contestants could simply give in and invest energy in finding another resource. On the other hand he could choose to fight. However, fighting is costly. In fact if both contestants decide to fight, they are both worse off than had they both decided not to fight and give in. Contestant A then has the following ordering over the possible results of their conflict: A prefers most the situation in which she gets the resource and B just gives in and moves away. Second on A's ordering is the outcome in which they both give in and share the resource. A costly fight has been avoided. Third is the situation in which B gets the resource after A has given in. Again a fight has been avoided and A has enough energy left to move on and try and find another resource. Utterly disastrous would be the situation in which both decide to fight, thus draining each other's energy to get any resource. Since B's ordering is symmetrical to that of A, we can conclude that they are involved in a *chicken game* (see figure 1).³

		B	
		give in	fight
A	give in	2 2	3 1
	fight	1 3	0 0

Figure 1, the chicken game

³ This name is derived from a game of bravado apparently played by some American teenagers Rapoport, Anatole (1967), "Exploiter, Leader, Hero, and Martyr: The Four Archetypes of the 2 x 2 Game", *Behavioral Science* 12: 81-84.. Biologists know this game as the *Hawk-Dove* game Smith, John Maynard (1982), *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.. I follow the social scientist's nomenclature here.

What should A and B respectively choose? There is no dominant strategy for either player. Their best reply depends upon the reply of the other. There are three equilibria in this game: two equilibria in pure strategies and one mixed equilibrium. Which one will be selected, if any?

To answer this question, the following assumption is crucial. We assume, as is typical for evolutionary game theory, that this is not a unique situation. There are many occasions where people like A and B will be in conflict over scarce resources. In other words, the game is a repeated game, in which a finite set of players is anonymously, randomly paired in each round.

Suppose that both players know exactly those things that classical game theory presupposes. That is, they know the possible outcomes and each other's payoffs. Furthermore, they know the other to be rational and all this is common knowledge. However, any additional information about the role of the contestants, even the labels of the strategies (i.e., "fight" and "give in") is unknown to them. One can demonstrate that in that case the only viable equilibrium in the long run is the mixed equilibrium.

Suppose that originally the population exists of "fighters" only. The average pay-off to a typical fighter equals 0: the result of an encounter with another fighter. Suppose that a mutant peace-lover enters the population. The average pay-off of this latter is much higher than that of the fighters. A peace-lover would end up with 1 whereas the typical fighter gets 0. In other words, there is an evolutionary pressure on the member of this population to switch from fighting to giving in. What is more, the more peace-lovers there are the higher the chances that two peace-lovers will meet, which would result in 2. So the push towards the peace-loving strategy will become stronger as more and more players switch strategy. However, this pressure is not such that the entire population will evolve towards peace-loving behavior. For if the entire population were to consist of peace-lovers, a mutant fighter would do extremely well, scoring 3 on all his encounters, which would generate a trend away from the peace-loving strategy.

This does not mean that this population would cycle between a state near all fighters and a state of nearly all peace-lovers. We can make the analysis of this game a bit more precise. There is a so-called mixed equilibrium between those two states, which is that situation where the a player fights in each encounter with probability p .⁴ A strategy is in equilibrium if and only if the

⁴ It is a difficult question how to interpret the mixed equilibria in standard game theory. This difficulty has been recognized since the beginnings of game theory. See for example Luce, R. Duncan and Howard Raiffa (1957), *Games and Decisions: Introduction and Critical Survey*, New

equilibrium strategy I does at least as well against itself as any mutant (i.e. a strategy other than I), that is $E(I, I) \geq E(J, I)$.⁵ Let I be the equilibrium strategy that fights with probability p , and J a mutant strategy that fights with probability q , then $E(I, I) \geq E(J, I)$ if and only if $p=0.5$. Note that following strategy I gives a typical player in this population an average pay-off of 1.5.⁶

More important that this result is the fact that this equilibrium is *evolutionary stable*. The concept of evolutionary stable strategy (ESS) was first introduced by the mathematical biologist John Maynard Smith:

An ESS is such a strategy that, if all members of the population adopt it, then no mutant could invade the population under the influence of natural selection.⁷

This stability is the case when two conditions are fulfilled. Let I and J be strategies, pure or mixed. Then I is an ESS if and only if:

(1) For all J it is the case that $E(I, I) \geq E(J, I)$, and

(2) For all J , such $I \neq J$, either $E(I, I) > E(J, I)$ or $E(I, J) > E(J, J)$

Of these two conditions (1) is the equilibrium condition (“ I is a best reply to itself”) and (2) is the stability criterion. Condition (2) requires either that I is the unique best reply against itself, or that it does better against other strategies than those strategies against themselves. In either case, mutants will not be able to invade the population.

York: Wiley.. On the one hand, one could think of them as a description of the composition of the population. This is the favored interpretation among evolutionary game-theorists, social scientists and biologists. Alternatively one can regard the mixed equilibrium as a description of the behavior of each individual player – as I do here. Each player fights in 50% of the encounters. There are hard methodological and philosophical questions to be put to either interpretation. I criticize the evolutionary interpretation in [verbeek, forthcoming #848].

⁵ Here we see one of the problems associated with the evolutionary interpretation of strategies as strategy profiles of a population. In order to calculate whether the system is in equilibrium we have to imagine that an *average* representative of each type of population is matched against the other. But, as I just stipulated, if we interpret the strategy pairs as descriptions of the evolutionary state of a population, it is hard to imagine which interpretation we can give to such matching.

⁶ Note also that the opponents of I will receive 1.5 regardless of the strategy they follow. This follows from the fact that the equilibrium strategy is a mixed strategy. Mixed strategies are “equalizer strategies”.

⁷ Smith, John Maynard (1982), *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.. Smith is a bit sloppy in his formulation here. Strictly speaking, a strategy is an ESS not only if no single mutant could invade the population, but also if that is not possible for very small groups of mutants. In the formal definition, this is taken into account.

The equilibrium strategy that fights with $p=0.5$ satisfies these two conditions. We have already demonstrated that (1) holds, so now we only need to verify that (2) holds. Since $E(I,I)=E(J,I)$, I is an ESS if and only if $E(I,J)>E(J,J)$. That is, the mixed equilibrium is stable if and only if it does better against mutants than the mutants against themselves. Let J fight with probability q , then $E(J,J)=3+q-q^2$. $E(I,J)=3.5-q$ which is larger than $E(J,J)$ for any value of q .⁸ The stability of I means that the contestants will actually fight in 25% of their encounters. In such a situation there is no norm to respect someone's property worthy of that name. It is not so far fetched to stipulate that a society that is described by the parameters of this model is in a state of nature, that is, a state in which there are no norms.

So far, we have assumed that the agents in this model are stripped of any knowledge except for what classic game theory allows them. That is, they do not know anything of the situation, other than the payoffs in the matrix. Does that matter? Modern evolutionary game theory has shown convincingly that it does. Perhaps the most elegant way of bringing this out is given by Skyrms (1996) and [Vanderschraaf, 1995 #846], though the idea is older.⁹ Suppose both contestants were to observe something a signal that is strictly speaking exogenous to the game. For example, suppose they were to observe who has possession of the resource before the encounter. Then they could adopt a contingency strategy that would lead them to one of the equilibria in pure strategies. For example, they could play the strategy "if possessor fight, if challenger give in".

If the population were to adopt such a strategy each individual will do better than under the symmetrical, mixed equilibrium. What is more, in the chicken game the asymmetrical strategy pairs each are an ESS and the mixed strategy is not. To see this, let p_A be the probability with which a player chooses

⁸ A less formal way to elucidate the stability of the mixed equilibrium is the following. Assume that a population of 100 individuals is in equilibrium when half the population fights and half the population gives in. Suppose that one person in this population switches strategy, the chance that the other person will fight in the next encounter is no longer 0.5, but 0.51. In that case giving in becomes marginally more attractive than fighting, so the pressure is on the fighters to give in until the equilibrium is established once again.

⁹ It can be found in Lewis, David (1969), *Convention: A Philosophical Study*, Cambridge: Harvard University Press.

Smith, John Maynard (1982), *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

Sugden, Robert (1986), *The Economics of Rights, Co-operation and Welfare*, Oxford: Basil Blackwell.. The insight that an ESS in a game with more than one ESS can be selected as result of an correlated equilibrium is relatively new. It has several advantages to think of the selection of the ESS in this way, most notably, it allows for the evolution of meaningful signals in a state of nature. See also [Vanderschraaf, 1995 #846].

to give in when challenger, and p_B be the probability with this player chooses to give in when possessor. A strategy in the game of figure 1 can be described as the pair (p_A, p_B) . Suppose I is the asymmetrical strategy which plays “When challenger give in; when possessor fight”, or in our notation, $(1,0)$.¹⁰ Let $J=(q_A, q_B)$ be a mutant strategy, with $q_A \neq 1$ and $q_B \neq 0$. Finally, let us assume that the probability that a player is possessor equals 0.5. In other words, one has an even chance to being possessor (or rather, *first* possessor) as one has chance to be challenger.¹¹

Substituting the values of figure 1, we can calculate $E(I,I)$ and $E(J,I)$:

$$(3) \ E(I,I) = 0.5[E(I_A, I_B) + E(I_B, I_A)] = 2 \text{ (where } I_A \text{ and } I_B \text{ stand for following strategy I when A and when B respectively)}$$

$$(4) \ E(J,I) = 0.5[E(J_A, I_B) + E(J_B, I_A)] = 0.5[q_A + 2q_B + 3(1-q_B)] = 1.5 + 0.5q_A - 0.5q_B$$

To meet the equilibrium condition (1) it has to be the case that $E(I,I) \geq E(J,I)$. This inequality holds if and only if $0.5 \geq 0.5q_A - 0.5q_B$, which is necessarily true. So I is an equilibrium strategy. Since $I \neq J$, it follows that $q_A \neq 1$ and $q_B \neq 0$. But then $E(I,I) > E(J,I)$ and the stability condition (2) is satisfied. Therefore, I is also stable.¹²

What this means is that in a population of I-players a deviation from that regularity immediately generates “punishment” for the deviant. Not because others react negatively to the deviation, but because it generates a bad result in comparison to what could have been had the rule been followed. Such regularities then are, as Sugden (1986) puts it, *self-enforcing*. Should an agent deviate from the norm, others in the population have no incentive to deviate as well if the equilibrium is stable.

One can also see how such a stable strategy would emerge in a population that previously did not follow it. In the chicken game, suppose some members of the population deviate from this mixed strategy, i.e., no longer give in with $p=0.5$, but adopt the strategy “Give in when challenger, otherwise fight.” An

¹⁰ Note that the mixed equilibrium then can be described by the probability pair $(0.5, 0.5)$.

¹¹ This last assumption is not necessary for the proof of stability, however it facilitates the arithmetic considerably.

¹² Note that the alternative strategy $K=(0,1)$ is also stable. However, in a population consisting mainly of I-followers it would do very poorly indeed. This is precisely the point of the stability criterion. All that stability implies is that new arrivals in a population consisting of agents following a stable strategy are most successful if they also follow that strategy.

individual deviant agent has no reason to rethink her decision in the next round. No matter how many in the population will keep on playing the mixed strategy, she herself is doing at least as well as the others who continue to play the mixed strategy. However, if she happens to encounter one of the other agents who play the same asymmetrical strategy, these two are able to generate better results for themselves than the rest of the population. As we have seen above, the expected payoff for this agent is 1.5 if she plays against an agent who follows the mixed equilibrium strategy. But should she encounter a fellow deviating agent who follows the same strategy as she does, she will receive an expected payoff of 2 on average. Thus, one can expect that other individuals, once they learn which signal the deviant strategy is using for her asymmetrical strategy, will adopt that strategy as well.¹³

Stability is an important feature in the analysis of norms. It corresponds to the second of our criteria for the existence of a norm, that is, deviations from the regularity are possible without the regularity necessarily collapsing in the subsequent rounds. Stability, therefore, is an interesting feature to look for. Indeed, it plays a central role in the definition of conventions. The strategy of “Giving in when challenger, otherwise fight” is a *convention*, which is defined as any stable equilibrium in a game that has two or more equilibria.

“Giving in when challenger, otherwise fight” is the accepted convention since there are more possible equilibria in the distribution of scarce resources. Other signals than possession could be used. Moreover, the convention could equally well be “giving in when possessor, otherwise fight”.

§4 Some preliminary remarks on the model

Returning to the central characteristics of conventionalism discussed in section 1, we can now see why evolutionary game theoretical analysis of norms is a form of conventionalism. To think of social norms as conventions in this technical sense means that one accepts the thesis of conformity with regards to them. Once one sees that if, for example, “The first possessor owns a particular resources” is the accepted convention in a particular community, this implies that it is rational to observe that right precisely because (almost) everybody else

¹³ In [Verbeek, forthcoming #848] I utter some reservations about the way one should think of the learning of the conventional strategy. In particular, I do not believe that we can model “social evolution” after the way in which biological evolution operates.

observes it. Therefore, conformity explains why reductionism is true for the norm to respect property.

The truth of reductionism does not lead one to be skeptical with respect to the normative force of norms. A standard criticism of the view that it is rational to comply with norms is that it leads to a dilemma. Either it is rational to do what the rule prescribes, but then it is rational for reasons other than the rule (e.g., because it is utility maximizing to do so). Or these other reasons do not obtain, but then it is hard to see why it is rational to comply with the norm. Therefore, either the rule is superfluous, or compliance is irrational. The conventionalist analysis of norms avoids this dilemma. It shows that one can be rational in genuinely complying with a norm (that is, one of several possible stable equilibria) precisely because the norm requires one to act in a certain way.

Finally, one should be careful when making straightforward inferences from these results to actual rules of property. The whole set of rules we refer to as the norms of property – if you will, the *institution* of property – is much more complex than this analysis of “first seizure” norms might lead one to conclude. For example are the rules for the transfer of property also stable equilibria in a chicken game? Having said that, it does not mean that I am a skeptic about the whole enterprise of showing that the norms of property are conventions. However, I doubt that all norms of the institution of property can be interpreted as stable equilibria in chicken games.

§5 Deviation from the norm

Having discussed the model at length, we should return to our original question whether evolutionary game theory is successful as an analysis of norms. We have seen that norms should be understood as more or less spontaneous patterns of regularity in the behavior of members of a group (characteristic (1)). We have seen that these regularities can spread in a population through an evolutionary process, which we could interpret as learning (characteristic (2)).¹⁴ We have seen, moreover, that the stability of the norms can be traced back to the nature of the norms themselves; that is,

¹⁴ Actually, I have several objections to this claim. I am unconvinced that we can think of learning as a process of replicator dynamics, at least not as this process is usually understood. See also [Verbeek, forthcoming #848].

conventions are self-enforcing. This self-enforcing character is the result of the fact that agents following the norm stand to benefit more than deviants. In other words, it is rational to comply with norms because they have a conventional character.

This is as far as evolutionary game theory can bring us towards a deeper understanding of norms. In order to account for the subsequent defining characteristics of norms we will have to introduce additional assumptions. Some of these are at odds with evolutionary game theory.

The third of the criteria presented in section 1 is that deviation from the norm is possible. In other words, agents can break the norm. So if norms are to be understood as conventions in the technical sense in which it is used here, we need to know if deviation from the convention is possible.

On the face of it, this seems an anomaly within the conventionalist model. After all, the whole analysis of conformity to social norms is based on the notion that it is conducive to the agent's goals and preferences to conform. Conventions are supposed to be self-enforcing. So if social norms are to be understood as conventions, conformity with them is straightforwardly rational.

This poses the theory with a dilemma. Either deviation from the equilibrium that forms the convention is sometimes rational or it is not. If the former is the case, the convention is not an equilibrium. According to the definition of an ESS given in section 3, if strategy I is an ESS, it must be a best reply against itself. In other words, I must be an equilibrium. Now if there are alternative strategies that do better than I in a population consisting mainly of I-players, I is not an equilibrium. So if norms are conventions defined in terms of the ESS criterion, then it cannot be rational to deviate from them. Moreover, if the conventional rule is not an equilibrium it could not have emerged in the first place. For what could be the incentive for a rational agent to change strategies and start playing I?

The other horn of the dilemma is the following. If it turns out that rational agents will never deviate from the convention, the case for emergence can still be made as well for the conventional character of the regularity in question. But then we are stuck with the conclusion that conventions cannot be norms since an essential ingredient of the characterization of norms in section 3 is that deviation is possible. Something has to give here: either the possibility of deviation can be accommodated in the model, or it should disappear from the list presented in section 3.

The obvious way in which the possibility of deviation could be accommodated in the model while avoiding the dilemma is to argue that

deviation could occur but that it is irrational. However, that does not seem to be true of all actual existing norms.¹⁵ Many modern conventionalists are aware of this. For example, Sugden (1986) mentions three reasons why people would deviate from the norm. First, people make mistakes. Conventions have to be learned. Human agents may fail to grasp the strategies involved or they may interpret them incorrectly (i.e., unconventionally). Secondly, Sugden argues people sometimes are weak-willed and they break the norm because of their lack of will power. People sometimes are not able to withstand the temptation to act in a way they know to be detrimental to their own interests. Just like the first explanation, this one avoids the dilemma by recognizing that deviation is irrational.¹⁶

However, Sugden also believes there are situations in which it is in one's interest not to follow the norm. For example, in the chicken game an agent, Boris, being extremely strong, might decide that he is willing to challenge a small-built possessor and fight him. The costs Boris will incur might be considerably less than the value of the resource to him in this specific case. A plausible conventionalist theory of norms has to admit that considerations such as these might lead to deviations from the established convention. By the same token, such a theory has to give an explanation of these deviations. As we have seen, the possibility of rational deviations challenges the very point of the conventionalist enterprise. For here we have genuine instances where deviation from an established convention is beneficial.

Basically, there are two ways to incorporate these considerations in the model. First, we could assume that norm following agents are to some extent characterized by moral motives, or cooperative virtues, as I will call them. These motives prevent them from deviating under such circumstances. Alternatively, one could assume that norms are backed up by sanctions. The

¹⁵ One example is the rule that one ought to have a valid ticket in public transport in my hometown of Amsterdam.

¹⁶ In [Verbeek, forthcoming #848] I argue that the assumption of weakness of the will does not fit well within the idea that conventions spread through social evolution, because the genetic algorithms that are used to model this evolutionary process do not assume conscious rationality or knowledge on behalf of the player. Instead players are modeled as mimicking or copying each other's behavior. This makes sense in a biological framework, but seems hard to defend in a social context.

Another complication is that, in order for such a thing as weakness of the will to be possible, one has to make assumptions about the psychology of agents, about the structure of their will, e.g., along the lines sketched by Frankfurt, Harry G. (1971), "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68: 5-20. or Watson, Gary (1975), "Free Agency", *Ibid.* 72: 205-220.. If one therefore wants to work out the notion of weakness of will within the framework of the theory of rational choice as it is applied in the conventionalist analysis, one needs a more elaborate framework than is presently available. This is not an easy task, but an attempt is made in [Verbeek, forthcoming #848].

deterrent effect of sanctions prevents deviance where this otherwise might be beneficial.¹⁷

§6 Sanctions

The typical reaction to deviation from a norm comes in the form of a sanction. The received view is that norms need sanctions to be stable. If sanctions were too light, or indeed absent, people would not obey social norms and consequently, such norms would disappear. Hobbes has often been interpreted as subscribing to this view. Such a view is, on the face of it, diametrically opposed to the central insight of the conventionalist project. Whereas the former claims the need for a system of sanctions, the latter argues that norms do not need sanctions for their stability since they are self-enforcing.

Is there room for sanctions in our model of property norms? In section 2, I claimed that one of the criteria of a norm is that deviation is met with criticism and that this criticism is a reason for the deviant to revise his or her behavior. The first question is why an otherwise rational agent would punish the past deviant actions of another agent. On the face of it, punishment seems pointless since the harm has been done already. Besides there are quite a number of costs associated with the practice of punishing. First, detection of deviant behavior often involves costs. Police departments all over the world spend a considerable amount of resources on criminal investigations. Parents spend long tedious discussions with their children to find out who took the cookies from the jar. The time such discussions take could have been spent on more pleasurable activities.

Secondly, sanctioning itself can be very costly. Imprisonment in a standard facility costs society more (on average per inmate) than the rent of a modest apartment. Informal sanctions can also be very costly. In most big cities there is an incentive to “look the other way” when a crime is being committed, especially if it is a violent one. After all, if you intervene the perpetrator might turn on you instead.

Being punished is usually costly for the deviant. In general the loss of freedom due to a prison sentence is considerable and is valued negatively. Similarly with informal sanctions: to be ostracized by one’s fellows is a sanction most people seek to avoid. Therefore, we can assume that sanctions can be a

¹⁷ Of course, these two assumptions are not incompatible.

reason to avoid deviance. Whether this is because of its deterrent effect on the potential deviant, or because of the assurance to third parties, I will not discuss here.

Against the costs of punishment there is only one possible benefit for the agent who sanctions. Other possible deviants might be persuaded not to break the convention in the future and the deviant in question might be convinced to revise his behavior as a result of all the sanctions. But this benefit typically extends to other members of society as well. If my punishing you deters you from future deviation, it is not only me who benefits but also everybody else with whom you will interact in the future. So while we can understand why criticism, in the form of costly punishments, can be a reason for the deviant to revise his or her behavior (namely because of its deterrent effect on future behavior), we still have a problem regarding the question why one would punish a deviant given that it is costly to do so.

I refrain from giving these comments a more formal treatment here.¹⁸ It turns out that even in a simple coordination game, in which there is a minimal degree of conflict of interests, the detection of deviance and the production of sanctions is problematic. In fact, it turns out that a system of sanctions is itself a collective good; a collective good that presupposes norms for its production, distribution, and application.

In general, there are two kinds of interaction problems involved in these norms of sanctions. First, there are coordination problems. The question is who will administer the sanctions. That remains a real problem even if the population has all the characteristics of a so-called privileged group, i.e., it is the case that for at least some individuals the costs of detection and punishment are outweighed by the benefits of less (or no) deviations from the convention in the future.¹⁹ Secondly, there are serious n-person prisoner's dilemmas. Given that it is costly to detect and punish deviants, there is always the temptation not to contribute. The more so since the benefits of anybody's contribution to the system of sanctions fall equally on every member of the population. In other words, even if it is clear who has to contribute, there is still the problem that for those agents individually not contributing dominates contributing.

¹⁸ In [Verbeek, forthcoming #848] I give an elaborate analysis of the problems related to the production of sanctions.

¹⁹ For the notion of privileged groups, see Olson, Mancur (1965), *The Logic of Collective Action*, Cambridge: Harvard University Press. See also Hampton, Jean (1987), "Free Rider Problems in the Production of Collective Goods", *Economics and Philosophy* 3: 245-273..

§7 Sanctions and cooperative virtues

What has this analysis taught us? In the first place, it has become clear that the detection of deviance and the production of punishment is not necessarily a rational reaction to the violation of a convention. The pattern that emerges here is the following. Once we allow for sanctions to be part of what makes a convention into a norm (and indeed we have good reasons to do so), we find that the production of those sanctions already presupposes the existence of a norm prescribing sanctioning in such cases. Those norms should also be understood as conventions in the sense of evolutionary game theory.

Following Kavka (1983), I will refer to such norms as higher-order norms. They are norms regulating the coordination and production of the sanctions associated with the deviance from other, lower-order social norms. Given the analysis in section 5, it should be possible that otherwise rational agents sometimes deviate from those conventions. The next question is “What happens if a higher-order norm is violated?” If we hold on to the notion that an essential ingredient of a norm is that deviance is met with sanctions, then deviance from a higher-order norm should also be met with sanctions. However, given that a system of sanctions is itself a collective good, there is a need for yet higher-order norms to regulate the production of those sanctions. In other words, we have started a regress where every norm is backed up by an infinite set of higher-order norms regulating the sanctioning.

This poses a problem for the conventionalist project. The analysis of deviation we gave in section 5 showed that a certain degree of violation of a norm can be expected given the sources of deviance that Sugden mentions. But if it is true that there are infinitely many higher-order norms connected to any first-order norm, there will be a substantial amount of norm violations. Suppose somebody violates a first-order norm. Then there is a second-order norm that prescribes punishment. This norm can be broken, in which case there is a third-order norm prescribing punishment. This norm can be broken, and so on. What this means is that the possibility of violating any first-order norm brings with it the possibility of more norm violations.

To see this, suppose that the rate of deviance with regards to a certain first-order norm in a population is q , where $0 < q < 1$. Suppose that this rate is constant with respect to all the higher-order norms related to this one. That would mean that the chance of norm deviance with regards to this first-order norm or any of its higher-order norms is: $q + q^2 + q^3 + q^4 + q^5 + \dots = \frac{q}{1 - q}$. This

is always larger than q itself. That means in fact that violation of a first-order norm brings with it more violations of other (higher-order) norms. It also implies that, as long as the rate of deviance is not too large, i.e., $q < 0.5$, the chance that there will be any violation of the first-order norm or any of its higher-order norms is smaller than 1. There is no certainty of deviation.

Is this increase of the chance of violations a problem? Evolutionary game theory argues that it is not. In the limit, there will be no more deviance. The claim is that at each higher level the chance of deviance is smaller. What is the chance that at the n th level the norm to punish those violating the $n-1$ norm will be broken? In order for the norm on the n th level to be applicable, violations up to the $n-1$ level have to have occurred. This chance then is exactly q^n . Given that $0 \leq q < 1$, this chance approaches 0 as n increases. In other words, at some sufficiently higher level, the norms will no longer be broken. Insofar as a norm depends on sanctions for its stability (and we have seen that it does not do so exclusively), the regress is not a problem.

However, the regress could be vicious. As we have seen there is only so much deviation that a convention can withstand. Once the number of deviations becomes too large, for whatever reason, it pays to switch strategies as well. This holds for first-order norms, but of course also for higher-order norms. So suppose that a convention of the n th-order to punish non-punishers of the $n-1$ order changes as result of the all too frequent noncompliance (for whatever reason). That means that the convention to punish those not participating in the system of sanctions of the $n-1$ order is subject to pressure to change as well. And so on all the way down. Therefore, the more elaborate the system of sanctions related to a certain first-order norm, the less stable the first-order norm is. In other words, the regress can be a vicious one. The regress gains in viciousness if we introduce loops in the chain of punishment. If it is possible that an agent will be called upon to punish negligence to punish negligence to punish. . . himself, there is even less incentive to punish.

Why would this vicious regress be a problem? Well, it is a problem to the extent a system of social norms depends on the presence of sanctions to ensure compliance. So we need to break the chain of higher-order sanctions somewhere. It is at this point that cooperative virtues play an important role. The idea is that somewhere in the chain the agents should be disposed to comply with the norms applicable to them.

For example, if normal citizens break the law, judges punish them. Judges are bound by norms in their jurisdiction and their discretion. They are judged by higher judges should they break norms or not apply them properly, and so

all the way up to the Supreme Court. On this level however, we cannot expect the judges to occasionally break the law, even if that would be in their own interests. We can and have to expect that they will uphold the law according to the norms laid down for their conduct. In other words, this is a level where the cooperative virtues need to be internalized to a large extent. A plausible theory of norms therefore needs to presuppose the existence of such dispositions.

§8 Resentment

If the analysis of conventions is correct and the possibility of deviating from them is allowed, this seems to imply that conventions are best understood as rules of thumb. They would be sources of advice on how to act in standard situations, allowing her to act otherwise in certain admittedly exceptional situations. In general it is best to stick to the rule, but a situation could come up where it is not rational to do so and if one realizes this, one should not conform to the rule.²⁰ Also, as we have seen in the previous sections, there is nothing about conventions that implies that agents *necessarily* react in a negative way to deviations. That just depends on the effect sanctions will have, as well as the several other factors identified in section 7. Hence the charge that we are to understand conventions as rules of thumb.²¹

However, this is not how we typically think of the norms that regulate social life. In other words, the question now before us is whether the analysis of norms as conventions falls short of the mark. It cannot account for the last three characteristics of norms, which refer to the so-called “internal” aspect of norms. Indeed if norms are indeed conventions we need to establish that conventions have some moral force. Some authors have proposed that a psychological mechanism gives conventions this force.²² This is the mechanism of *resentment*. People resent deviant behavior. More precisely, people resent it if others behave in a manner that is contrary to their expectations and this resentment is a spontaneous reaction to deviant behavior.²³

²⁰ Note that if this were all that there is to conventionalism, it would be very similar to a form of indirect utilitarianism.

²¹ For a related worry, see Sugden (1986, 149).

²² Most notably, Pettit, Philip (1990), “Virtus Normativa: Rational Choice Perspectives”, *Ethics* 100: 725-755.

Sugden, Robert (1986), *The Economics of Rights, Co-operation and Welfare*, Oxford: Basil Blackwell.

²³ Pettit, Philip (1990), “Virtus Normativa: Rational Choice Perspectives”, *Ethics* 100: 725-755. argues that rational choice theorists have focused on the costs of sanctions too much. He

The spontaneous and emotional aspect that is associated with the term 'resentment' is very fortunate. For what we are looking for are not the type of well-calculated choices rational agents make in deciding how to react to somebody who does not behave as they expected. Rather we are looking for the type of plain anger that deviation arouses in most people. This reaction is not necessarily a rational reaction (i.e., instrumental in the pursuit of one's goals, whatever they are) on a traditional understanding of rationality. We are, in effect, leaving the framework of analysis discussed in section 6.²⁴

In order to understand how it is possible that such resentful reactions cause deviating agents to revise their behavior, it has to be the case that the expectations of other agents do matter to deviants. Moreover, the anticipation of such resentment is a reason not to deviate from the established convention. Resentment should have a deterrent effect. If the way in which resentment causes deviant agents to revise their behavior is a rational response, it is unclear why it is so. On one reading, we can interpret the assumption that the expectations of others matter to us as postulating a desire that (most) agents in fact have. This is the interpretation Sugden and Pettit endorse.²⁵

From now on I will refer to both the spontaneous reaction of resentment and the assumed propensity of the expectations of others to motivate us as the *mechanism of resentment*. If one accepts that there is such a mechanism, it seems that we have a model that accounts for all aspects of norms. We have an explanation why, in general, people obey norms and we have an explanation why the occasional deviation is met with negative reactions that cause deviants to revise their behavior.²⁶ In short, it seems that we have a complete analysis of norms. If such a mechanism is present, we have an account of norms as conventions that seems to cover almost all the elements listed in section 3. What

stresses that we should replace what he calls the *interest-based* explanation of norms with an *attitude-based* explanation of norms. The point of the latter is that punishment is not necessarily costly since it may be a spontaneous reaction to deviance. Detection need not cost anything either since a sanction can be the spontaneous reaction to deviance when confronted with it. Pettit is referring to such things as the spontaneous criticisms of bystanders or victims when confronted by deviance and other forms of sanctions with an informal character. Such an attitude-based explanation is exactly what Sugden is proposing here.

²⁴ Therefore, I disagree with Pettit that such spontaneous resentful reactions are without cost. Rather, the associated costs are irrelevant to the resenting agent.

²⁵ In the next section, I defend an alternative understanding of the role of resentment and why it can have an impact on human agents. People do not conform to the expectations of others for fear their behavior might be resented. Rather, people feel resentment because they have reasons to expect compliant behavior other than the potential resentment.

²⁶ Sugden (1986) and Pettit (1990) are not the only authors who assume that something like the mechanism of resentment is "in the head" of norm-obeying agents. Elster, Jon (1990), "Norms of Revenge", *Ethics* 100: 826-885. and Frank, Robert (1988), *Passions within Reason*, London: W. W. Norton & Company, Inc. share these ideas.

is more, it looks like we can avoid the introduction of moral motives into the analysis altogether. Whatever more formal sanctions cannot bring about is achieved through the mechanism of resentment.

§9 Moral force and resentment

There is still, however, a deficiency in the conventionalist analysis of social norms. It is this deficiency I seek to expose in this section. This deficiency will throw an alternative light on the mechanism of resentment and its proper role. What is lacking is an analysis of the last of the aspects of norms listed in section 1, to wit, the requirement that the negative reaction to deviant behavior is regarded as reasonable, as fitting, by the agents concerned. This requirement is not accounted for in the mechanism of resentment. An analysis of resentment that does offer criteria for this reasonability necessitates us to assume that norm following agents are characterized by moral motives.

Resentment involves the following aspects: (1) The resentment concerns other agents and their actions; (2) resentment involves that the interests of the one who resents are at stake and the one who is resented knows this to be the case; (3) resentment implies mutual expectations about each other's behavior; and (4) these expectations presuppose moral motives of the kind that I will refer to as cooperative virtues.²⁷

Condition (1) is easily verified. One does not feel resentment against the outburst of Mount Pinatubo volcano, no matter how much damage it has done or the degree in which one's expectations about the likelihood of the event were reasonable.²⁸ Conditions (1) and (2) imply that resentment is a backward-looking response. It follows after someone's actions have damaged one's interests. This means that resentment is to be distinguished from negative reactions that aim at correcting the deviant in such a way that she will conform on future occasions. In the latter type of actions, future conforming behavior is intended. In the case of resentment there is no such intention. All that the agent intends to do is to express her indignation to the deviant. Resentment focuses on antecedent events. This antecedence is conceptual, not necessarily temporal, in its nature. If one anticipates that one's interests will be damaged for no good

²⁷ I am agnostic about the question whether these conditions are necessary, sufficient or neither. I simply assume that they tell us something about the phenomenon of resentment

²⁸ Except of course when one lives in an animistic culture where one's worldview allows for personification of such things as volcanoes.

reason, there seems to be a good cause for resentment even though the actual harm has not yet occurred.

Damage to one's interests should not be interpreted simply as damage directly to one's own interests. Some people resent breaches of conventions which damage the interests of others. For example, the outrage people feel when they hear of a drunk who, by driving on the left, has caused a serious accident. They were not directly influenced. However, one could argue that they were indirectly concerned. The presence of people not following the convention is a threat to my interests even when I am not directly involved. I might meet them sometime in a similar situation. Also it might be a threat to the continuation of the convention. If some people no longer stop at red traffic lights this could induce a collapse of the rule to stop for red traffic lights. This presumably is detrimental to my interests as a user of roads. Traffic will be less safe. So interests of the one who resents are at stake whether directly or indirectly.²⁹

It could be argued that resentment is itself arbitrary. People feel resentment concerning all kinds of things. The situation in former Yugoslavia is cause for general outrage, as is the outcome of the war of 1812, the institution of slavery as it occurred more than hundred years ago, and so on. In none of these situations can it be argued that the interests of the agents who feel this resentment are at stake. I would hesitate to refer to it as resentment, but rather reserve the term *indignation* for it. Resentment, as I understand it, refers to a very specific class of cases. It concerns the damage to one's interest or to those of others within the group in which the particular convention that is violated is established.³⁰

The second clause of condition (2) – i.e. that the resented knows the interests of others are at stake – more or less implies condition (1). Actions involve intentions. If an act is an act of hurting somebody's interests this means that the actor intends this effect, or is aware that this is a side effect of his action. If not, it is not an act of hurting at all. For example, if I hurt you unintentionally you may feel anger that I did not know, or blame me for not knowing, but you do not resent me simply because of this harming action. A

²⁹ Another defense of this line of argument invokes the third condition: this driver *exploits* other peoples' compliance with the traffic rules. Because others stick to the rules the risk to him of drunken-driving has become acceptable in spite of the consequences to the other participants in the traffic. This driver then acted *unfairly*.

³⁰ Of course one could argue that the international community has its own norms, international laws, that are within the scope of conventionalism just as more local norms are. Then, in my terminology, the reaction of people to the atrocities of the Serbs in Bosnia is an example of resentment, since the interests of the international community are at stake.

very common form of excuse for harm done is “I am sorry. . . , I did not realize. . . .” The fact that this functions as an excuse, as a statement to avoid unjustified resentment, shows also that knowledge of the interests at stake is a crucial precondition to resentment.

The third condition introduces the mutual expectations of the actors concerned. These mutual expectations typically are (based on) conventions. Resentment, therefore, follows deviance from a convention. In what follows I will show, by using an example, why the phenomenon of resentment involves mutual behavioral expectations. It will become clear why these expectations presuppose the moral motives of fairness and trustworthiness (the fourth requirement).

Suppose there are two entrepreneurs, A and B, who compete on the local market for pins. One way of competition, legally open to them, would be to start selling pins for less than the cost of production, in order to gain a substantial share of the market. If both were to do so simultaneously, they would lose a lot of money. If just one were to do that for a short period, this entrepreneur would increase his share of the market substantively, while his competitor would almost be out of business. In other words, these two are engaged in a repeated prisoners’ dilemma. Suppose that they agree informally not to start a price war thus ensuring their mutual stable profits. A and B both rely on the norms laid down in this agreement.³¹ Both expect that the other will honor the agreement. Moreover, each expects that the other expects that they expect, and so on, that they will honor the agreement. In other words, the code laid down in the agreement becomes a convention in their market behavior.

Suppose that A at one point does not honor the agreement because it is especially advantageous for him to renege at that point. He dumps a large amount of pins under the cost price. B resents this. He will resent this because it was reasonable for him to expect that A would not violate the agreement between them. Now if A had reneged on the agreement from the beginning, B would not have come to believe that A would honor the agreement. Nor would B believe that A believes that B believes that A would honor the agreement, etc. The course of events would not have warranted the belief that A would honor the agreement. In that case, B might be sad that they did not succeed in coordinating their pricing policies, or deplore A’s shortsightedness, but he would not resent what A did. B resents that A reneges and gains advantage

³¹ Or rather, A and B rely on the norms specified in the contract against the background of an established convention that prescribes honoring one’s agreements.

knowing that B would honor the agreement. In other words, A exploits B's preparedness to honor the agreement. This preparedness, as we have seen, is based on the existence of the mutual expectations about honoring the agreement. The presence of such beliefs makes B's resentment of A's behavior reasonable and intelligible in the first place. This illustrates that resentment presupposes the existence of mutual expectations as stated in condition (3).

This is different from the way Sugden and Pettit believe resentment works. For them, B could be said to resent A regardless of the existence of those previous mutual expectations. If A had made it clear at the time of the agreement that he would renege on the agreement as soon as it would be in his interest to do so, B's reaction – no matter how unreasonable – would still be a matter of resentment. Moreover, it would be a reaction that would matter to A, regardless that it is totally unreasonable. It is clear from the example that I find their explanation implausible.

What would make it reasonable for B to expect that A would indeed honor the agreement? After all, we stipulated that it was an informal contract with none of the legal sanctions applicable to enforce it. In other words, why did B trust A to honor the contract? Several reasons spring to mind. First, B might count on A's rationality, reasoning that it is in A's interest to honor the agreement as this is the best way to maximize his profits. However, that is not real trust. What is more, the example shows that it is probably not true that it is in A's interest here and now to uphold his end of the agreement. These are exactly the instances in which we are interested in the discussion of the role of resentment in the conventionalist theory. Resentment comes into the picture precisely in those situations where it is rational to deviate from the convention.

A second, possible and more important reason why it is reasonable for B to trust A is B's belief that A is trustworthy. He believes that A will not let him down even when it occasionally might be advantageous to do so, not for some external reason, but because B trusts him. Since A will hold similar beliefs about B, their respective trustworthiness uphold their mutual expectations. Trust, then, is important to understand why the expectations about each other's behaviors based on the convention are reasonable, even when it is occasionally rational not to live up to those expectations. Trust assumes trustworthiness, or at least the expectation of it; therefore condition (4) is fulfilled as well.³² Where

³² I discuss the relation between trust and trustworthiness at some length in [Verbeek, forthcoming #848].

there is resentment, moral motives (e.g., trustworthiness) help explain why the expectations that form the background of the resentment are reasonable.

What is more, because of these moral motives we can understand that the reasonability of one's expectations in itself is not enough to explain resentment. These motives are necessary to understand why our expectations are reasonable to begin with, but they also explain something else. Consider again B's reaction to A's reneging. B is not just stating his surprise that A did not live up to his reasonable expectations. Nor is he urging A to stick to the agreement the next time since that would serve both their competitive interests best. Resentment is more than that. It is the outrage that one feels because the other "has let me down"; the anger that follows the realization that the other "has exploited me". This passionate reaction is understandable against the background of the sort of moral motives I described.

It might be helpful to point out once more the differences and similarities between the view of Sugden and Pettit and mine on this point. Whereas Sugden assumes that resentment is the result of a psychological disposition that reacts to falsified expectations, my view is that resentment is a reasonable reaction in so far as these expectations themselves are reasonable. Sugden assumes the existence of another psychological mechanism, such that the expectations of others matter to us. On my view it is not resentment that explains why people would feel the moral force of a norm even when it is in their interest to deviate here and now. People feel the moral force of norms precisely because they see that the expectations of others about their behavior based on this norm are reasonable. In order to explain why this is reasonable (and not just rational), one needs to point at certain moral motives, cooperative virtues, such as trustworthiness. Sugden thinks he can do without these.

This analysis of resentment and its relation to norms and moral motives is not restricted to pin manufacturers such as A and B. It is fundamental to all social norms whether they are traffic regulations, norms of taxation or conventions of reciprocity. Insofar as it is possible that sometimes it is in the interest of the agent to deviate, we have to suppose that such deviations usually are met with resentment and that this resentment is intelligible as such only against the background of the four conditions I mentioned.

This is not to say that each and every agent following a norm is morally disposed in such a manner. What is being supposed is that each agent expects that others are, in general, trustworthy, fair; in short, that others are morally motivated in following a norm. Moreover, such expectations are reciprocal. These expectations make resentment of deviant behavior reasonable. Finally,

the existence of such expectations will stand up to rational scrutiny if and only if there is a general tendency that agents have such moral motives. In other words, the mutual expectations about each other's trustworthiness have to be founded on something. This foundation is supplied by moral motives. Most people who interact with each other according to social norms are morally motivated.

What does this prove? We have seen that resentment is characterized by certain conditions. These conditions contain the possibility and the presence of a cooperative virtue, e.g., trustworthiness. This also explains why other people's expectations matter to us even when the resentment of others is all the "punishment" we receive for not living up to reasonable expectations. Ideally rational actors probably might be indifferent to this "punishment", yet most flesh-and-blood people are not. This means that people are likely to feel emotions varying from uneasiness to guilt when they break a norm, i.e., when they do not honor the reasonable, justifiable expectations of others.

§10 Conclusions

We are now in a position to see that evolutionary game theoretical explanations of norms have to assume moral motives. We have seen at several places within the theory that cooperative virtues are necessary. First, in those situations in which it is be rational to deviate from the norm. The second place where we saw the need for cooperative virtues emerging was in the context of sanctions. The third place where cooperative virtues prop up in the theory is in the discussion of resentment. We saw that we need to have cooperative virtues to explain, first, why in some cases it is reasonable to suppose people will conform to the convention. Secondly, cooperative virtues are necessary to come to grips with the often passionate character of resentment in response to norm violations. Resentment is not an expression of surprise, nor is it urging the other to conform the next time. Rather resentment is the emotional response to unfair or untrustworthy behavior, a response to "being let down" by (the) other(s).

In any group or society where social norms are stable, the members of that group are characterized by a certain degree of cooperative virtuousness. Does this mean that game theory is unimportant for a plausible theory of social norms? Should we instead focus on the appropriate psychology of norm

following agents? I do not think so. Evolutionary game theory has earned its place in textbooks in social science and philosophy. I do not know of a single alternative theoretical model that comes so close to the mark when it comes down to formulating a complete and plausible (for conventionalist) theory the nature of social norms. However, I hope it is clear that we are not there yet.

References

- Bartsch, Renate (1983), "Normconflicten en de Wijzen van het 'Zijn' van Normen", *Filosofie en Praktijk* 4: 114-133.
- Binmore, Ken (1994), *Playing Fair (Game Theory and the Social Contract; vol 1)*, Cambridge: The MIT Press.
- Binmore, Kenneth (1992), *Fun and Games. A Text on Game Theory*, Lexington, Massachusetts: D. C. Heath and Company.
- Elster, Jon (1990), "Norms of Revenge", *Ethics* 100: 826-885.
- Frank, Robert (1988), *Passions within Reason*, London: W. W. Norton & Company, Inc.
- Frankfurt, Harry G. (1971), "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68: 5-20.
- Hampton, Jean (1987), "Free Rider Problems in the Production of Collective Goods", *Economics and Philosophy* 3: 245-273.
- Hart, Herbert L. A. (1961), *The Concept of Law*, Oxford: Oxford University Press.
- Kavka, Gregory (1983), "Rule by Fear", *Noûs* 17: 601-620.
- Lewis, David (1969), *Convention: A Philosophical Study*, Cambridge: Harvard University Press.
- Luce, R. Duncan and Howard Raiffa (1957), *Games and Decisions: Introduction and Critical Survey*, New York: Wiley.
- Olson, Mancur (1965), *The Logic of Collective Action*, Cambridge: Harvard University Press.
- Pettit, Philip (1990), "Virtus Normativa: Rational Choice Perspectives", *Ethics* 100: 725-755.
- Rapoport, Anatole (1967), "Exploiter, Leader, Hero, and Martyr: The Four Archetypes of the 2 x 2 Game", *Behavioral Science* 12: 81-84.
- Schelling, Thomas (1960), *The Strategy of Conflict*, Cambridge: Harvard University Press.
- Skyrms, Brian (1996), *Evolution of the Social Contract*, Cambridge: Cambridge University Press.
- Smith, John Maynard (1982), *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

- Sugden, Robert (1986), *The Economics of Rights, Co-operation and Welfare*, Oxford: Basil Blackwell.
- Watson, Gary (1975), "Free Agency", *Journal of Philosophy* 72: 205-220.
- Winch, Peter (1990), *The Idea of a Social Science and its Relation to Philosophy*, London: Routledge & Kegan Paul.