CHAPTER 1

INTRODUCTION

1.1 Norms

The ubiquity of norms is overwhelming. There are (detailed) norms regulating our behavior in the community at large, norms that regulate our actions in the schools 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. There are norms regulating spoken language, as well as our communications on electronic mail and on paper. The sequence of characters on this page is dictated by a norm. There are detailed norms guiding our behavior in traffic. 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. Our lives are pervaded by norms of all kinds. Some of these norms are rules we have set only for ourselves. They determine our individual actions and habits. For example, I have made it a rule never to leave home without my keys. However, many norms regulate the interactions between people. These norms are such that we expect each other to observe them. We believe that others expect the same of us. By these characteristics, these norms can be identified as social norms.

The question that motivated me to write this study is a simple one. Why should one obey social norms? What reasons can there be to comply with the multitude of formal and informal social norms that structure our lives? One important reason is that compliance is rational. Many of these norms are helpful in coordinating our actions and ensuring, if not optimal, at least, satisfying results from our efforts. The rules of the English language ensure that people following these rules can exchange information. The norms laid down in the traffic rules ensure an acceptable level of safety for all participants when going from A to B. Compliance to such norms seems rational given the goals people usually have. However, this is not quite as self-evident in the case of other social norms. For example, the norm to pay taxes is useful from the point of view of society as a whole. But when the annual tax assessment drops in one's mailbox many people are at least

tempted to avoid paying the full assessment. So even though compliance with some norms is rational, this does not seem to be the case for other norms.

1.2 Rationality in action

Rationality is a term philosophers as well as social scientists like to use. All too often it is not entirely clear what is meant by it. In order to prevent similar confusion, I will give the rough outlines of the concept of rationality that is used throughout this work.

Rationality is a term denoting the appropriateness of an action given the goals of the agent. Therefore, "rationality" is used in the minimal sense, to wit, as instrumental rationality. This notion does not judge the goals agents pursue in their actions. The theory is neutral about these. The focus of this notion of rationality is on the choice of action against the background of given goals and evaluations of the agent. This is the dominant sense in which it is used in many of the social sciences, particularly in economics, as well as in many areas of philosophy. This notion of rationality is the most basic. Whatever else one subsumes under the heading of rationality in action, instrumental rationality is part of it.[1] Instrumental rationality is first and foremost a normative notion. It expresses a judgment on the appropriateness of the means chosen to pursue the goal. Alternatively, it can be used as a hypothesis about the behavior of people. The rationality of action then is an empirical question.

The above mentioned usages of the notion are indeterminate as long as one does not spell out the implications that follow from it. In other words, one needs a theory of rationality in order to apply it. The most prevalent theory, the theory of rational choice, implements these ideas as follows.[2] Agents face situations in which they must choose from a number of possible actions or "strategies".[3] Each action leads to an outcome. Which outcome will be realized is not exactly foreseen by the agent since it depends on the exact state of the world, which is usually unknown to the agent. However, it is supposed that the agent attaches a certain probability to the realization of each possible outcome. It is further supposed that the agent has a preference-ordering over each of these possible outcomes. The ordering then can be represented with the use of a function that assigns numbers to the set of outcomes. This is the utility function. If the preferences of the agent satisfy certain requirements, a utility function can be constructed that

---

[1] Nozick (1993).

[2] Excellent introductions in the basic ideas of rational choice theory, as well as its aims can be found in Resnik (1987) and Binmore (1992).

[3] That is, they cannot abstain from acting. Or, alternatively, abstention is an action as well.

has the following two characteristics.[4] First, of any two outcomes, the one that is most preferred is the one with the highest utility value. Secondly, this function is such that to each action an expected utility can be attributed. The expected utility of an action is the sum of the utilities of the possible outcomes of the action weighted with the corresponding probabilities. The implication is that the agent should choose or chooses the action with the greatest expected utility. The other implication is that the agent should make her preferences consistent in the way these requirements specify.

Things become more complicated when we consider the problems agents face when the outcomes of an action depend in part on the actions of others. This is the domain of game theory. The basic insight of this part of rational choice theory is that the agent needs to consider in her deliberation about what to do, the deliberations of others as well. For example, in a simple game of *rock, paper, scissors*, the agent, in her choice, needs to take into consideration the possible choices of the other.[5] Suppose she believes her opponent will choose "rock". She then should choose "paper" since "paper covers rock". However, she should also realize that her opponent can replicate the same reasoning, which would lead him to choose "scissors" since "scissors cut paper". But then she would have done better to choose "rock" as "rock breaks scissors", which gives her opponent a reason to choose "rock" and she is back where she started. The most fundamental result of modern game theory is the idea that even in such situations there is an equilibrium of choices, that is, that choice which is the best reply to the best reply of the other(s). This has led game theoreticians to identify such puzzles as the prisoners' dilemma, where the unique equilibrium is such that all players receive their one but worst outcome while at least one Pareto-superior outcome is available. For this reason, the strategies in the prisoners' dilemma have been labeled as *cooperation* and *defection*.[6] The prisoners' dilemma shows that defection is uniquely rational for all, even though all parties would do better if they would cooperate. The prisoners' dilemma turned out to reveal the structure of some of the basic problems with regard to compliance with norms.[7]

---

[4] A simple, but precise demonstration of these requirements and the resulting function can be found in Luce and Raiffa (1957, ch. 2). More elaborate constructions are those of Savage (1954) and Jeffrey (1965). Resnik (1987) summarizes their findings.

[5] This is a game where two players are to announce their choice of any one of these three items simultaneously. The rules imply that there is an intransitive order over the three actions, such that rock beats scissors, scissors beat paper, and paper beats rock.

[6] I assume that the reader is familiar with the prisoner's dilemma and the conventional notation of games in matrices. Binmore (1992) is an excellent introduction An informal introduction is Dixit and Nalebuff (1993). For the little game theory I use in this work, Dixit and Nalebuf's book is more than adequate.

[7] This is to such an extent at one point, that Axelrod (1984) referred to the prisoners' dilemma as "the *e-coli* of the social sciences".

1.3 Norms and rationality

As stated above, compliance with some norms seems rational, assuming the usual preferences people have. Compliance with other norms seems irrational. Particularly in situations which can be modeled as prisoners' dilemmas. The taxpayer example of section 1 is such a dilemma. Though it is better for all, if everybody would pay taxes, it is better for the individual not to pay taxes while everybody else does. Moreover, if the individual expects that others will not pay taxes, she avoids the worst outcome by evading her taxes as well. Since this reasoning is relevant to each individual in the population, it would seem that compliance is irrational for all.

This is not necessarily the case. As we shall see, the so-called conventionalist analysis of norms aims to demonstrate that rationality prescribes compliance for all social norms. However, this analysis falls short of its goal. It turns out that the stability of social norms presupposes a certain attitude of the rule followers. They need to possess the cooperative virtues in some degree. These virtues are qualities, such as trust and fairness, which ensure that people will cooperate in the one-shot prisoners' dilemma, as long as they can expect that others will do so as well.

The conclusion is that the cooperative virtues, in conjunction with rationality, justify compliance with social norms. However, these two – the cooperative virtues and rationality – are at odds, at least on the standard theory of rational choice. This is the central problem of this study. If the existence of social norms requires that people in general be trustworthy and fair, how does this relate to the desirability of having such motives as an individual? In other words, should an otherwise rational agent be glad to be morally disposed or should she try to get rid of that ballast on the first occasion? What we are asking for then is a rational foundation for the cooperative virtues.

Before presenting the plan of this work, let me state its most important conclusion. The discrepancy between the cooperative virtues and the prescriptions of instrumental rationality is real. However, it is not as deep a divide as is commonly assumed. The reason for this is that the standard theory of rational choice that prescribes defection in all one-shot prisoners' dilemmas is mistaken. In more technical terms this study aims to show the following: conditional cooperation – conditional, that is, upon the cooperative actions of the other – in a-synchronous two-person prisoners' dilemmas where one player chooses her strategy after the other, is rational. This does not cover all cooperative choices made by trustworthy and fair agents, but it drastically reduces the accusation of irrationality. It shows that in many instances of the prisoners' dilemma there are virtues, rational virtues, in cooperation.

1.4 The plan of this work

Chapter two deals with a particular, radical, theory of social norms: the conventionalist theory. Conventionalism entails the endorsement of two propositions. First, that it is rational to comply with social norms. Secondly, that part of the reason why this is rational is because it is known that all, or a sufficiently great number of others, comply as well. The aim of the chapter is to make the case for the necessity of cooperative virtues for the emergence, stability of and compliance to social norms.

In general, a social norm exists in a group when (a) there is a certain regularity in the behavior of the members of the group; (b) deviations of this regularity are possible; (c) these deviations can be recognized by the members of the group; (d) the regular behavior can be learned; (e) deviations of the regularity are criticized; (f) this criticism usually is a reason to revise behavior; (g) the criticism is justified, correct.

The convention theory of Robert Sugden provides a model that can account for (a) up to (d).[8] According to the theory, a social norm is a rule deviation from which is not in the agent's interests. The motive of the actor to comply to the norm is her (self-)interest given the expectations of the actions of others. However, this motive does not support compliance in all situations to which the theory is supposed to apply. From this I conclude that in such situations an extra motive, other than (self-)interest, is necessary to account for compliance to norms.

A separate section is dedicated to the question whether fear of sanctions can explain compliance. I show that this is not a plausible solution. First, because it does not apply to those situations where an agent could deviate from the norm undetected. Secondly, because this argument amounts to a regress. Sanctions are costly to administer. Norms are needed which specify who should sanction what behavior and how. These norms themselves should be backed up by sanctions, which implies that there should be norms regulating those sanctions, and so on. In order to understand the effectiveness of a system of sanctions, an appeal to cooperative virtues is necessary. This conclusion is strengthened by the observation that norms often are accompanied by certain reactive attitudes such as resentment, guilt, and shame. A closer analysis of resentment reveals that it refers to the existence of cooperative virtues.

In chapters three and four, I undertake a discussion of the content of the cooperative virtues. The possible dispositions discussed in chapter two are those that lead to unconditional cooperation. In particular, altruism and one kind of "process-oriented preferences", i.e., Kantianism, are discussed. It will appear that dispositions for unconditional cooperation cannot be part of the cooperative virtues.

In chapter four, I look at dispositions for conditional cooperation. Reciprocal cooperation, trust, and fairness are discussed. All these

---

[8] Sugden (1986).