1

The Rules We Live By

Introduction

Despite the ubiquitous reference to the concept of social norms in the social sciences, there is no consensus about the power of social norms to direct human action. For some, norms have a central and regular influence on human behavior, while for others, the concept is too vague, and the evidence we have about norm compliance is too contradictory to support the claim that they appreciably affect behavior. Those who doubt that norms have a behavior-guiding force argue that human behavior only occasionally conforms with the dominant social norms. If the same norms are in place when behavior is norm-consistent as when it is norm inconsistent, why should we believe that norms mediated any of it?

Much of the discussion about the power norms have to affect behavior arises from a confusion about what is meant by 'norm.' A norm can be formal or informal, personal or collective, descriptive of what most people do, or prescriptive of behavior. In the same social setting, conformity to these different kinds of norms stems from a variety of motivations and produces distinct, sometimes even opposing, behavioral patterns. Take for example a culture in which many individuals have strong personal norms that prohibit corrupt practices and in which there are legal norms against bribing public officers, yet bribing is widespread and tolerated. Suppose we were able to independently assess whether an individual has a personal norm against corruption. Can we predict whether a person, who we know condemns corruption, will bribe a public officer when given a chance? Probably not, but we could come closer to a good prediction if we knew certain factors and cues are present in this situation and have

an influence on the decision. The theories of norms we have inherited, mainly from sociology, offer little help, because they did not develop an understanding of the conditions under which individuals are likely to follow a norm or, when several norms may apply, what makes one of them focal.

A first step in the direction of a deeper understanding of what motivates us to follow a norm is to clarify what we mean by a social norm. 'Norm' is a term used to refer to a variety of behaviors, and accompanying expectations. These should not be lumped together, on pain of missing some important features that are of great help in understanding phenomena such as variance in norm compliance. Inconsistent conformity, for example, is to be expected with certain types of norms, but not with others. In this chapter I put forth a 'constructivist' theory of norms, one that explains norms in terms of the expectations and preferences of those who follow them. My view is that the very existence of a social norm depends on a sufficient number of people believing that it exists and pertains to a given type of situation, and expecting that enough other people are following it in those kinds of situations. Given the right kind of expectations, people will have conditional preferences for obeying a norm, meaning that preferences will be conditional on having expectations about other people's conformity. Such expectations and preferences will result in collective behaviors that further confirm the existence of the norm in the eyes of its followers.

Expectations and conditional preferences are the building blocks of several social constructs, though, not just social norms. *Descriptive norms* such as fashions and fads are also based on expectations of conformity and conditional preferences, and so are *conventions*, such as signaling systems, rules of etiquette, and traffic rules. In both cases, the preference for conformity does not clash with self-interest, especially if we define it in purely material terms.[1] One can model descriptive norms and conventions as solutions to coordination games. Such games capture the structure of situations where there exist several possible equilibria and, although we might like one of them best, what we most want is to coordinate with others on *any* equilibrium; hence we act in conformity to what we expect others to do. Descriptive norms and conventions are thus representable as equilibria of original coordination games. *Social norms*, on the contrary, often go against narrow self-interest, as when we are

[1] What one most prefers in these cases is to 'do as others do,' or to coordinate with others' choices.

required to cooperate, reciprocate, act fairly, or do anything that may involve some material cost or the forgoing of some benefit. The kinds of problems that social norms are meant to solve differ from the coordination problems that conventions and descriptive norms 'solve.' We need social norms in all those situations in which there is conflict of interest but also a potential for joint gain. The games that social norms solve are called mixed-motive games.[2] Such mixed-motive games are not games of coordination to start with, but social norms, as I shall argue, *transform* mixed-motive games into coordination ones. This transformation, however, hinges on each individual expecting enough other people to follow the norm, too. If this expectation is violated, an individual will revert to playing the original game and to behaving 'selfishly.' This chapter thus starts with a precise definition of social norms and only later considers what differentiates such norms from descriptive norms and conventions. Because all three are based on expectations and conditional preferences, I pay special attention to the nature of expectations (empirical and/or normative) that support each construct.

The definition of social norm I am proposing should be taken as a *rational reconstruction* of what a social norm is, not a faithful descriptive account of the real beliefs and preferences people have or of the way in which they in fact deliberate. Such a reconstruction, however, will have to be reliable in that it must be possible to extract meaningful, testable predictions from it. This is one of the tasks I undertake in Chapters 3 and 4. An important claim I make in this chapter is that the belief/desire model of choice that is the core of my rational reconstruction of social norms does not commit us to avow that we always engage in conscious deliberation to decide whether to follow a norm. We may follow a norm automatically and thoughtlessly and yet still be able to explain our action in terms of beliefs and desires.

The simplistic, common view that we conform to norms either because of external sanctions or because they have been internalized flies in the face of much evidence that people sometimes obey norms even in the absence of any obvious incentive structure or personal commitment to what the norm stands for (Cialdini et al. 1990). Many who postulate internal or external incentives as the sole reasons for compliance also maintain compliance is the result of a conscious process of balancing costs

[2] Well-known examples of mixed-motive games that can be 'solved' (or better, 'transformed') by norms of fairness, reciprocity, promise-keeping, etc., are the Prisoner's Dilemma, the Trust game, and Ultimatum games.

and benefits, culminating in a decision to conform or to transgress. Yet personal experience tells us that compliance is often automatic and unreflective: Even important social norms like those that regulate fair exchanges and reciprocation are often acted on without much thought to (or awareness of) their personal or social consequences. Whereas the literature on social norms has traditionally stressed the deliberational side of conformity, in this book I want to emphasize its automatic component. Both aspects are important, but too much emphasis on conscious deliberation may miss crucial links between decision heuristics and norms, as I explain in this chapter and the next.

Whenever we enter any environment, we have to decide how to behave. There are two ways to reach a decision. One is somewhat ideally depicted by the traditional rational choice model: We may systematically assess the situation, gather information, list and evaluate the possible consequences of different actions, assess the probability of each consequence occurring, and then calculate the expected utility of the alternative courses of action and choose one that maximizes our expected utility. I dub this the *deliberational* route to behavior. The process of rational deliberation ending in the choice of a course of action is likely to be costly in time, resources, and effort and to require considerable skill. The deliberational way to behavior is likely to be chosen when one is held accountable for one's choice; when the consequences may be particularly important and long-lasting; or when one has the time, knowledge, and disposition to ponder over alternative choices. But even in these cases deliberation may fall short of the ideal. Behavioral decision theorists have gathered compelling evidence that actors systematically violate the assumptions of rational choice theory (Camerer 2003). Thus the deliberational way need not assume perfect rationality. It only requires conscious deliberation and balancing of what one perceives (or misperceives) as the costs and benefits of alternative courses of action. On occasion we do engage in conscious deliberation, even if the process is marred by mistakes of judgment and calculation.

A second way to reach a decision relies on following behavioral rules that prescribe a particular course of action for the situation (or a class of similar situations). These guides to behavior include habits, roles, and, of course, norms. Once one adopts a behavioral rule, one follows it without the conscious and systematic assessment of the situation performed in deliberation. The question of how a particular behavioral rule is primed is of great interest. The answer is likely to lie in the interplay of (external) situational cues and (internal) categorization processes. These processes

lie beyond awareness and probably occur in split seconds. Models of mental processes (Lamberts and Shanks 1997) suggest that, when faced with a new situation, we immediately search for cues about how to interpret it or what is appropriate behavior for that situation. It is conjectured that we compare the situation we face with others we remember that possess similar characteristics, and that this comparison activates behavior that is considered most "normal" for this type of situation. The comparison process is one of 'categorization,' of finding relevant similarities between the current context and other ones we have experienced in the past. To efficiently search our memory and group a new event with previously encountered ones, we use cognitive shortcuts. Cognitive shortcuts play a crucial role in categorization and the subsequent activation of scripts and schemata.[3] Consequently, they are responsible for some norms rather than others being activated in different situations. Let us call this route to behavior the *heuristic* route. In the heuristic route, behavior is guided by *default rules* stored in memory that are cued by contextual stimuli. Norms are one class of default rules. According to the heuristic route, norm compliance is an automatic response to situational cues that focus our attention on a particular norm, rather than a conscious decision to give priority to normative considerations. On the heuristic view, norms are context-dependent, meaning that different social norms will be activated, or appear appropriate, depending on how a situation is understood. In turn, our understanding of a situation is influenced by which previous contexts we view as similar to the present one, and this process of assessing similarities and 'fitting' a situation into a pre-existing category will make specific norms salient. I spell out in detail the process of drawing social inferences and categorizing in the next chapter.

The distinction between deliberational and heuristic routes to behavior is a useful simplification, and it should be taken as such. The truth is that we often combine the two routes, and what is a staple of the heuristic process can also be an object of deliberation. Conformity to a norm, for example, is not always an automatic, nondeliberational affair. Especially when we are tempted to shirk an obligation, the thought of the personal and social consequences of alternative courses of action is often present and important in determining our choice. I want to stress, again, that deliberation is not synonymous with 'rational deliberation', in part

[3] Schemata are cognitive structures that contain knowledge about people, events, roles, etc. Schemata for events (e.g., a lecture, going to a restaurant, playing a chess game) are also called scripts. Chapter 2 further elaborates on the roles of scripts and schemata.

because the list of possible mistakes and cognitive impairments with which our decision processes are fraught is potentially very long. *Rational deliberation* is better conceived of as an ideal type, against which we measure the amplitude of our deviations. What is important in deliberation is the *conscious* processing of information and evaluation of options. Whether ideally or less than ideally rational, deliberation refers to beliefs and desires of which we are *aware*. Deliberation is the process of consciously choosing what we most desire according to our beliefs. In the deliberational view, beliefs and desires (preferences) are treated as mental states of which we are conscious, at least in the course of deciding which action to take.

The problem with taking beliefs and desires to be conscious mental states is that they can then play no role in the heuristic route to behavior. There is, however, a long and reputable philosophical tradition that takes beliefs and desires to be *dispositions* to act in a certain way in the appropriate circumstance. According to the dispositional account, to say that someone has a belief or a preference implies that we expect such motives to manifest themselves in the relevant circumstances. Thus, for example, one might automatically obey a norm of truth-telling without thinking of the beliefs and preferences that underlie one's behavior. These beliefs and preferences might become manifest only when they happen to be unfulfilled. To assess the nature of such beliefs and desires, all we need is a simple counterfactual exercise. Suppose we ask someone if he would keep telling the truth (as he normally and almost automatically does) in a world where he came to realize that people systematically lie. Our subject may answer in a variety of ways, but whatever course of action he claims he would choose, it is likely that he never thought of it before. *He did not know,* for example, that he would be ready to become a liar until he was put in the condition to reflect on it. Our subject may reason that it would be stupid on his part to keep telling the truth, as it would put him at an obvious disadvantage. Evidently his preference for sincerity is conditional on expecting reciprocity. If these expectations were not met, his preference would be different. Note that dispositions need not be stable: Preferences, for example, can be context-dependent, in the sense that even a small change of context may elicit different, even opposite, preferences. The research on framing effects shows just that (Tversky and Kahneman 1981). The heuristic way to behavior seems perfectly compatible with a dispositional account of beliefs and desires. Namely, the default rules that we tend to automatically follow are accompanied and supported by beliefs and desires that we become aware of only when they are challenged. Surprise in this case breeds awareness of

our underlying motives. Moreover, whenever a norm is 'cued' or made salient in a particular environment, the mechanism that primes it elicits the beliefs and preferences that support that particular norm. The remainder of this chapter presents a taxonomy of norms that relies on preferences and beliefs as 'building blocks.'

The idea that social norms may be cued, and hence manipulated, is attractive. It suggests that we may be able to induce pro-social behavior and maintain social order at low cost. Norms differ in different cultures, and what cues a Westerner into cooperation will probably differ from what cues a Mapuche Indian (Henrich 2000). In both cases, however, it may be possible to structure the environment in a way that produces desirable behavior. If you sail along the Italian coast, you will notice large beach posters that invite sailors not to litter and pollute "your" sea. In Sweden, instead, environmentalist appeals always refer to "our" environment. The individualistic Italians are seemingly thought to be more responsive to an invitation to protect a "private" good, whereas Swedes are expected to be sensitive to pleas for the common good. Knowing what makes people focus on the environment in a positive way can be a powerful tool in the hands of shrewd policymakers. Still, developing successful policies that rely on social norms presents several difficulties. To successfully manipulate social settings, we need to predict how people will interpret a given context, which cues will 'stand out' as salient, and how particular cues relate to certain norms. When multiple conflicting norms could apply, we should be able to tell which cues will favor one of them. Many norms are not socially beneficial, and once established they are difficult to eliminate. If we know what induces people to conform to "anti-social" norms, we may have a chance to curb destructive behavior. Without a better understanding of the mechanisms through which norms control our actions, however, there is little hope of predicting and thus influencing behavior. The mechanisms that induce conformity are very different for different kinds of norms. Consequently, a good understanding of their diversity will prevent us from focusing on the wrong type of norm in our efforts to induce pro-social behavior.

In the remainder of this chapter I will introduce the reader to my definition of social norms, descriptive norms, conventions, and the conditions under which one might see individuals following any of these. I shall especially focus on the four (individually) necessary and (jointly) sufficient conditions for a social norm to exist that I develop in the following pages: contingency, empirical expectations, normative expectations, and conditional preferences.

Social Norms

Social norms are frequently confused with codified rules, normative expectations, or recurrent, observable behavior. However, there are significant problems with such definitions of social norms. By the term *social norm*, I shall always refer to informal norms, as opposed to formal, codified norms such as legal rules. Social norms are, like legal ones, public and shared, but, unlike legal rules, which are supported by formal sanctions, social norms may not be enforced at all. When they are enforced, the sanctions are informal, as when the violation of a group norm brings about responses that range from gossip to open censure, ostracism, or dishonor for the transgressor. Some such norms may become part of our system of values, and we may feel a strong obligation to obey them. Guilt and remorse will accompany transgression, as much as the breach of a moral rule elicits painfully negative feelings in the offender. Social norms should also be distinguished from moral rules: As I shall argue in the following, expectations are crucial in sustaining the former but not necessarily the latter. In particular, conformity to a social norm is conditional on expectations about other people's behavior and/or beliefs. The feelings of shame and guilt that may accompany a transgression merely reinforce one's tendency to conform, but they are never the sole or the ultimate determinants of conformity. I will come back to this point later.

A norm cannot be simply identified with a recurrent, collective behavioral pattern. For one, norms can be either prescriptive or proscriptive: In the latter case, we usually do not observe the proscribed behavior. As anyone who has lived in a foreign country knows, learning proscriptive norms can be difficult and the learning process slow and fraught with misunderstandings and false steps. Often the legal system helps, in that many proscriptive norms are made explicit and supported by laws, but a host of socially relevant proscriptions such as "do not stare at someone you pass by" or "do not touch people you are not intimate with when you talk to them" are not codified and can only be learned by trial and error. In most cases in which a proscriptive norm is in place, we *do not* observe the behavior proscribed by the norm, and it is impossible to determine whether the absence of certain behaviors is due to a proscription or to something else, unless we assess people's beliefs and expectations. Furthermore, if we were to adopt a purely behavioral account of norms, nothing would distinguish shared fairness criteria from, say, the collective morning habit of brushing one's teeth. It would also be difficult to deal with those cases in which people pay lip service to the norm in public and deviate in private. Avoiding a purely behavioral account means focusing on the role expectations play in supporting those kinds of collective behaviors that we take to be norm-driven. After all, I brush my teeth whether or not I expect others to do the same, but I would not even try to ask for a salary proportionate to my education if I expected my co-workers to go by the rule of giving to each in proportion to seniority. There are also behaviors that can be explained only by the existence of norms, even if the behavior prescribed by the norm in question is never observed. In his study of the Ik, Turnbull (1972) reports that these starved hunter-gatherers tried hard to elude situations where their compliance with norms of reciprocity was expected. Thus they would go out of their way to avoid being in the role of gift-taker. A leaking roof would be repaired at night, so as to ward off offers to help and future obligations to repay the favor. Hunting was a solitary and furtive activity, so as to escape the obligation to share one's bounty with anyone encountered along the way. Much of the Ik's behavior can be explained as a successful attempt at *eluding* existing reciprocity norms. The Ik seemed to have collective beliefs about what sort of behavior was prescribed/proscribed in a given social context but acted in ways that prevented the underlying norms from being activated. Their practices demonstrate that it is not necessary to observe compliance to argue that a norm exists and affects behavior.

As Turnbull's example shows, having normative beliefs and expecting others to conform to a norm do not always result in a norm being activated. Nobody is violating the norm, but everybody is trying to avoid situations where they would have to follow it. Thus, simply focusing on norms as clusters of expectations might be as misleading as focusing only on the behavioral dimension, because there are many examples of discrepancies between normative expectations and behavior. Take the widely acknowledged norm of self-interest (Miller and Ratner 1998): It is remarkable to observe how often people (especially in the United States) expect others to act selfishly, even when they are prepared to act altruistically themselves. Studies show that people's willingness to give blood is not altered by monetary incentives, but typically those very people who are willing to donate blood for free expect others to donate blood only in the presence of a sufficient monetary reward (Wuthnow 1991). Similarly, when asked whether they would rent an apartment to an unmarried couple, all landlords interviewed in Oregon in the early 1970s answered positively, but they estimated that only 50% of other landlords would accept an unmarried couple as tenants (Dawes 1972). Such cases are rather common; what is puzzling is that people may expect a given norm to be upheld

in the absence of information about other people's conforming behavior and in the face of personal evidence to the contrary. Thus, simply focusing on people's expectations may tell us very little about collective behavior.

If a purely behavioral definition of norms is deficient, and one solely based on expectations is questionable, what are we left with? Norms refer to behavior, to actions over which people have control, and are supported by shared expectations about what should/should not be done in different types of social situations. Norms, however, cannot just be identified with observable behavior, nor can they be equated with normative beliefs, as normative beliefs may or may not result in appropriate actions. In what follows I introduce a definition of social norms that will be helpful in shedding light on the conceptual differences between different types of social rules. My definition coincides with ordinary usage in some respects but departs from that usage in others. Given the fact that the term has been put to multiple uses, it would be unrealistic to expect a single definition to agree with what each person using the term means. The goal of giving a specific definition is to single out what is fundamental to social norms, what differentiates them from other types of social constructs.

Besides helping in drawing a taxonomy of social rules, a successful definition should provide conditions under which normative beliefs can be expected to be consistent with behavior. This means that those conditions that are part of the definition of social norm would be used as premises in a practical argument whose conclusion is the decision to conform to a norm. This does not entail that we normally engage in such practical reasoning and deliberation and are consciously aware of our conforming choices. We should not confuse adopting a belief/desire explanatory framework with assuming awareness of our own mental processes. As I shall discuss in the last section, the fact that we are mostly unaware of our mental processes, and often are not fully conscious of what we are thinking and doing, is no objection to a belief/desire model of choice.

The definition I am proposing should be taken as a *rational reconstruction* of what a social norm is, not a description of the real preferences and beliefs people have or the way in which they in fact deliberate (if at all). The advantage of a rational reconstruction is that it substitutes a precise concept for an imprecise one, thus removing the conceptual difficulties and vagueness related to everyday usage. A rational reconstruction of the concept of norm specifies in which sense one may say that norms

are rational, or compliance with a norm is rational.[4] Not every rational reconstruction will do, though. For example, a rational reconstruction that is built on a belief/desire structure is constrained by the requirement that, were beliefs to be different (in a specified sense), we would expect behavior to change in predictable ways. In other words, a successful rational reconstruction must allow meaningful, interesting predictions to be made.

Conditions for a Social Norm to Exist

Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:

Contingency: i knows that a rule R exists and applies to situations of type S;
Conditional preference: i prefers to conform to R in situations of type S on the condition that:
(a) *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S;
and either
(b) *Normative expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
or
(b′) *Normative expectations with sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior.

A social norm R is *followed* by population P if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b′) are met for i and, as a result, i prefers to conform to R in situations of type S.

There are several features of the above definition that need explanation. First, note that a rule R can be a social norm for a population P even if it is not currently being followed by P. I defined P_{cf} as the set of 'conditional followers' of R, those individuals who know about R and have a conditional preference for conforming to R. I defined P_f as the set of 'followers' of R, those individuals who know about R and have a preference

[4] E. Ullmann-Margalit (1977) made one of the first attempts at explaining norms and norm compliance in a rational choice framework.

for conforming to R (because they believe that the conditions for their conditional preference are fulfilled). A behavioral rule R is a social norm if the set of its conditional followers is sufficiently large; a social norm is followed if the set of its followers is sufficiently large. Second, note that a social norm is defined relative to a population: A behavioral rule R can be a social norm for one population P and not for another population P'. Finally, the 'sufficiently large subset P_{cf} of P'' clause reflects the fact that social norms need not be universally conditionally preferred or even universally known about in order to exist. A certain amount of opportunistic transgression is to be expected whenever a norm conflicts with individuals' self-interest. The 'sufficiently large subset P_f of P_{cf}' clause reflects the fact that, even among conditional followers of a norm, some individuals may not follow the norm because their empirical and normative expectations have not been fulfilled. Moreover, even among the members of P_f, occasional deviance due to mistakes is to be expected. How much deviance is tolerable is an empirical matter and may vary with different norms. For example, we would expect P_{cf} (the proportion of conditional followers) to be equal to P in the case of group norms, especially when the group is fairly small, whereas P_{cf} will be close to P in the case of well-entrenched social norms. For new norms, or norms that are not deemed to be socially important, the subset P_{cf} could be significantly smaller than P. I will discuss deviance and its effects in later chapters, when I address the issue of norm dynamics. It should also be noted that I do not assume P_f (the proportion of actual followers) to be common knowledge. Different individuals will have different beliefs about the size of P_f and thus have different empirical expectations. If so, they will have different thresholds for what 'sufficiently large' means. What matters to actual conformity is that each individual in P_{cf} believes that her threshold has been reached or surpassed.

Condition 1, the *contingency condition*, says that actors are aware that a certain behavioral rule exists and applies to situations of type S. This collective awareness is constitutive of its very existence as a norm. Note that norms are understood to apply to classes or families of situations, not to every possible situation or context. A norm of revenge, for example, usually applies to members of a kinship group and is suspended in case of proven accidental death. A norm of reciprocity may not be expected to apply if the gift was a bribe, and the rules that govern fair allocation of bodily organs differ from those that regulate the fair allocation of university Ph.D. slots. Situational contingency explains why people sometimes try to manipulate norms by avoiding those situations to which the norm

applies (as the Ik did with food sharing and gift reciprocation) or by negotiating the meaning of a particular situation.

Condition 2(a), the *empirical expectations condition*, says that expectations of conformity matter. I take them to be *empirical expectations*, in the sense that one expects people to follow R in situations of type S because one has observed them to do just that over a long period of time. If the present situation is of type S, one can reasonably infer that, *ceteris paribus*, people will conform to R as they always did in the past. Notice that the fulfillment of Condition 2(a) entails that a social norm is *practiced* (or is *believed to be practiced*) in a given population (which may be as small as a group comprising a few members or as large as a nation); otherwise there would not be empirical expectations. Sometimes expectations are formed not by directly observing conforming behavior, but rather its consequences. This would happen, for example, with norms regulating private behavior. In this case, public support might be voiced for a norm that is seldom adhered to in private. If conformity to such a norm is believed to produce observable consequences, then observing such consequences will validate the norm. But if these consequences are the effect of other causes, people will draw the wrong inference and continue to believe that the norm is widely followed even when support is dwindling. Consider a norm of private behavior such as avoiding premarital sex; what we observe are the consequences of such behavior (teen pregnancy, etc.) or the lack thereof. If people take adequate precautions, there might be greater deviance than expected, but people might still believe that the norm is widely practiced in the population.[5] Norms regulating private behavior may thus present us with cases in which Conditions 2(a) and 2(b) are satisfied. However, as I shall make clear in discussing Conditions 2(b) and 2(b′), there are many individuals for whom 2(b′), the possibility of sanctions, is a *necessary* condition for compliance. Such individuals will believe they are expected to follow the norm but will not expect to be sanctioned for transgressing it [Condition 2(b′)], because deviance can be concealed. In this case, public endorsement of the norm may coexist with considerable private deviance.

The expectations mentioned in Condition 2(a) could, besides being empirical, also be *normative*, in the sense that people might think that

[5] I would venture the hypothesis that norms regulating private behavior may survive longer than other norms precisely because of the lack of direct observation of compliance. On the other hand, they may decay very quickly once the magnitude of deviance becomes public knowledge, as I discuss in Chapter 5.

everyone 'ought to' conform to *R* in situations of type *S*. The 'ought' implicit in a normative belief does not necessarily state an obligation. Take, for instance, a well-known convention such as the rule of driving on the right side of the road. We believe that people ought to follow that rule simply because, if they do not, they risk killing or being killed. If a person does not want to jeopardize her life, nor does she have an interest in causing harm to others, then we believe she 'ought to' follow the driving rule. The 'ought' in this case expresses prudential reasons and is akin to saying that, if you have goal *x* and the best available means to attain *x* is a course of action *y*, then you ought to adopt *y*. Consider, on the other hand, a rule of equal division. In this case, we may believe that others ought to 'divide the cake in equal parts' because this is the fair thing to do. We think they have an obligation to follow the rule, a duty to be fair. I do not ask for the moment what grounds this obligation, though I shall come back to this question later. At this point it is only important to make a distinction between a prudential 'ought' and the statement of an obligation. From now on, when I mention 'normative expectations' I will always refer to the latter meaning.

Normative expectations do not necessarily trump empirical ones, and very often they coexist. Many well-entrenched social norms are thought to be good or reasonable, and people often refer to these qualities in justifying their own compliance, as well as in expecting other people to comply. Yet there are also cases in which most people do not think that others ought to conform to a norm, even when they observe widespread conformity (i.e., the number of those prepared to sanction others is very small). This happens with norms that many, maybe most, people dislike and yet are followed by everyone. Wearing a veil may be an unpleasant requirement for many Muslim women, and they may not believe that one 'ought to' wear it (apart from prudential reasons). But if each woman holds the belief that she is expected to wear a veil, in the sense of believing that a sufficiently large number of people think she ought to wear a veil and prefer that she wears a veil (because it is her religious duty to do so), then she will feel great social pressure in that direction, and the result will be overall collective compliance. In this case the norm regulates public, observable behavior; hence a transgression is easily detected and likely to be punished. If it is not public knowledge that most women dislike the veil, a woman may even take widespread adherence to this norm as evidence that other women follow this practice out of a deep religious conviction, and infer that she is expected by everyone else to fulfill her religious duty as well. Everyone may secretly feel she is a deviant, but they

will never openly question the norm. I will discuss in Chapter 5 how such 'pluralistic ignorance' may be responsible for the survival of norms that most people dislike.[6] For now it is enough to emphasize that a normative interpretation of Condition 2(a) is not necessary for my argument.

Conditions 2(b) and 2(b′) tell us that people may have different reasons for conditionally preferring to follow a norm. Condition 2(b), the *normative expectations condition*, says that expectations are *believed to be reciprocal*. That is, not only do I expect others to conform, but I also believe they expect me to conform. What sort of belief is this? On the one hand, it might just be an empirical belief. If I have consistently followed *R* in situations of type *S* in the past, people may reasonably infer that, *ceteris paribus*, I will do the same in the future, and that is what I believe. On the other hand, it might be a normative belief: I believe a sufficiently large number of people think that I have an obligation to conform to *R* in the appropriate circumstances. For some individuals, the fulfillment of Conditions 2(a) and 2(b) is sufficient to induce a preference for conformity. That is, such individuals recognize the legitimacy of others' expectations and feel an obligation to fulfill them. For others, the possibility of sanctions is crucial to induce a preference for conformity. Condition 2(b′) says that I believe that those who expect me to conform also *prefer* me to conform, and might be prepared to sanction my behavior when they can observe it. Sanctions may be positive or negative. The possibility of sanctions may motivate some individuals to follow a norm, either out of fear of punishment or because of a desire to please and thus be rewarded. For others, sanctions are irrelevant, and a normative expectation is all they need. Condition 2(b′) does not say that transgressions *will* be punished and compliance rewarded. It only states that a sufficiently large subset of *P* *may* be capable and willing to sanction others. As we shall see in a moment, normative expectations are essential for the enforcement of social norms.

Now suppose Conditions 1, 2(a), and either 2(b) or 2(b′) hold. Each of them is a necessary condition for conformity to *R*, but *contingency*, *empirical*, and *normative expectations* are not jointly sufficient to produce conformity to rule *R* in situations of type *S*. I might expect others to follow a rule of equal division, and believe that I am expected to follow that rule

[6] What social psychologists call *pluralistic ignorance* is a psychological state characterized by the belief that one's private thoughts, attitudes, and feelings are different from those of others, when in fact they are not, in a situation where public behavior contradicting these private thoughts and attitudes is identical (Allport 1924; Miller and McFarland 1991).

too, but when it is my turn to 'cut the cake,' I may be tempted to get a larger share, especially if nobody is observing my action. If I do not, it must be that I *prefer* to conform to the rule. However, this is no simple, unconditional preference for conformity. Condition 2, the *conditional preference condition*, says this preference is *conditional* on expecting others to conform to *R* and either believing that one is expected to conform to *R* or believing that those who expect one to conform also have a preference for collective conformity and are prepared to punish or reward. If so, the counterfactual "If I were to believe that others do not follow R or do not expect me to follow R, then I would not want to conform to R" must be true. What I am saying suggests that following a social norm may be contrary to self-interest, especially if we define it in purely material terms. Thus it may be the case that, in the presence of monetary or otherwise 'material' rewards, I have a tendency to prefer more to less but will prefer to 'share' if I believe that I am in a situation in which some form of generosity is the norm, if I expect others to be generous, and if I believe them to think I 'ought to' be generous in the circumstances. In this case, I might prefer to behave generously. Note that the generous behavior induced by adherence to a norm should not be confused with other motives, such as altruism or benevolence.

Before we continue our discussion of Condition 2, let us look at an example that will hopefully clarify what I mean by saying that the motive to follow a norm should be distinguished from other motives. Consider playing a one-shot prisoner's dilemma, where C stands for Cooperate and D stands for Defect.

If the payoffs in Figure 1.1 represent sums of money, just by looking at them it is not obvious what a player will choose. Suppose Self, the row player, only cares about his 'material' self-interest and thus prefers DC to CC, CC to DD, and DD to CD. If B stands for best, S for second best,

Self \ Others	C	D
C	3, 3	0, 4
D	4, 0	1, 1

FIGURE 1.1. One-shot Prisoner's Dilemma

Self \ Others	C	D
C	S	W
D	B	T

FIGURE 1.2. One-shot Prisoner's Dilemma from the perspective of narrowly self-interested Self

T for third best, and W for worst, the preference ranking of a narrowly self-interested Self would look like that shown in Figure 1.2.

The narrowly self-interested person will always choose D, her dominant strategy. Self-interest, however, should not be confused with the desire for material incentives. A self-interested person is one whose ultimate desires are self-regarding, but these desires can involve 'immaterial' goods such as power and recognition, or the experience of 'benevolent' emotions. A self-interested person may want to 'feel good' (or reap social rewards like status and love) by reciprocating expected cooperation and in this case her preferences would look like those in Figure 1.3.[7]

Self \ Others	C	D
C	B	T
D	W	S

FIGURE 1.3. One-shot Prisoner's Dilemma from the perspective of benevolent Self

[7] I am assuming for simplicity that the benevolent individual is concerned with the material well-being of another. The same assumption holds for the pure altruist. However, their utility functions look very different. If x_i and x_j are the payoffs, respectively, of player i and player j, the pure altruist's utility will be $U_i f(x_j)$, and $\delta U_i/\delta x_j > 0$. The benevolent player's utility instead will be $U_i f(x_i, x_j)$, and the first partial derivatives of U_i with respect to x_i, x_j, will be strictly positive.

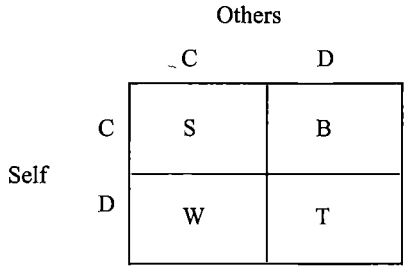


FIGURE 1.4. One-shot Prisoner's Dilemma from the perspective of altruistic Self

Note that a benevolent person would prefer CD to DC; that is, she would prefer, *ceteris paribus*, to be the righteous sucker rather than the spiteful cheat. This preference would probably be cost-sensitive, but if the costs are not too high, it makes sense to prefer to 'feel good about oneself' and be the loser rather than penalizing another to get some small benefit.

Benevolent motives are different from those of a pure altruist, whom I take to be a person whose ultimate desires are completely other-regarding. A pure altruist wants, first and foremost, the satisfaction of another's desires, at whatever cost to the self.[8] If the altruist believes his partner to be a narrowly self-interested type, the altruist's preference ranking would look like the one in Figure 1.4.

The person who instead follows a norm of generosity or cooperation need not have a desire to 'feel good': If the established norm is a cooperative one, provided Conditions 2(a) and either 2(b) or 2(b′) are met, the preference ranking of the norm follower will look like the one in Figure 1.5.

The norm follower's preferences are similar to those of the self-interested, benevolent person, with a crucial difference: For the benevolent person, it is better to be the 'sucker' than the 'crook' (CD is preferred to DC); but for the norm follower, the reverse may be true.[9] This distinction should not be interpreted as denying that individuals can be both benevolent and norm followers. Benevolence, however, is usually

[8] The choice to donate part of one's liver to an anonymous recipient is an example of altruism, because the risk of complications and even death from the procedure is sizable.

[9] Again, I am assuming for simplicity that the norm follower is not also benevolent. If this were the case, Figures 1.3 and 1.5 would coincide, at least in all those situations to which benevolence applies. In large, anonymous groups, where the effects of one's actions are insignificant, we may expect less cooperation (or not at all) from the benevolent person, whereas the norm follower would not be affected.

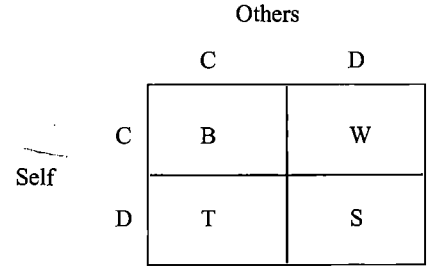


FIGURE 1.5. One-shot Prisoner's Dilemma from the perspective of norm-following Self

directed to people with whom we habitually interact and know well. As social distance increases, benevolence tends to decrease. If most people were benevolent toward strangers, we would need no pro-social norms of fairness, reciprocity, or cooperation. In particular, we would have no need for those norms that 'internalize' externalities created by behavior that imposes costs on other people. Thus it is plausible that one is guided by benevolence (or even altruism) in interacting with family and friends, but when interacting with strangers, be guided by social norms. Moreover, whereas benevolence toward those who are close to us should be a relatively stable disposition, generosity or cooperativeness with strangers will vary according to our expectations, as defined in Conditions 2(a) and 2(b) or 2(b′).

It may be objected that motivational distinctions are futile, because often observation cannot discriminate among them. If in a one-shot social dilemma experiment we observe consistent cooperative behavior, what can we say about the underlying preferences? If, as economists do, we take preferences to describe behavior and not motivation, what we observe is a 'revealed preference' for taking into account other people's welfare. *Why* we do that does not matter. Still, I believe motivations carry some weight. Up to now, most experiments have been geared to show that human behavior consistently deviates from the narrow, self-interested paradigm postulated by traditional economic models. Experiments have been very successful in this respect, yet they do not tell us why actors have other-regarding preferences. Is it altruism, benevolence, or are we priming norms of fairness and reciprocity? The answer is clearly important, and not just for the policymaker. What we now need is to test more sophisticated hypotheses about what goes on in the black box. To do so it is important to pay attention to the meanings of the concepts we use

(and test). To tell altruism and benevolence apart is not very difficult: If an altruist is informed that the other defected, the altruist should keep cooperating. Never mind there are very few such people around: If they exist, that is the way altruists will behave. The benevolent individual and the norm follower are more difficult to set apart. For one, a norm follower may also be motivated by benevolence. If, however, some norm followers are not benevolent, the distinction would be most clear in all those situations in which people are forced to choose between CD and DC. Suppose we identify a subset of people who 'conditionally cooperate' in one-shot Prisoner's Dilemmas. That is, controlling for their expectations, they cooperate whenever they expect others to cooperate, too. It should be possible to perform another experiment on the same individuals in which the only choice is one between being the sucker or the crook: The subject might be told that the other player will choose next, and will have to choose the opposite of what she does. Provided the personal cost is not too high, the benevolent person should prefer being the sucker. A person who instead followed a cooperative norm for reasons other than benevolence would see no reason to be the sucker (possibly provided the cost to the other person is not too great).

Condition 2 (the *conditional preference condition*) marks an important distinction between social and personal norms, whether they are habits or have moral force. Take the habit of brushing my teeth every morning. I find it sanitary, and I like the taste of mint toothpaste. Even if I came to realize that most people stopped brushing their teeth, I would continue to do so, because I have independent reasons for doing it. It is likewise with moral norms: I have good, independent reasons to avoid killing people I deeply dislike. Even if I were to find myself in a Hobbesian state of nature, without rules or rights, I would still feel repugnance and anguish at the idea of taking a life. With this I do not mean to suggest that moral norms are a world apart from other rules. Instead, by their very nature, moral norms demand (at least in principle) an unconditional commitment.[10]

[10] It might be argued that even what we usually understand as moral norms are conditional. One may be thoroughly committed to respect the sanctity of human life, but there are circumstances in which one's commitment would waver. Imagine finding oneself in a community where violence and murder are daily occurrences, expected and condoned by most. One would probably at first resist violence, then react to it, and finally act it out oneself. Guilt and remorse would in time be replaced by complacency, as one might come to feel the act of murder to be entirely necessary and justified. The testimonies of survivors of concentration camps, as well as the personal recollections of SS officers, are frightening examples of how fragile our most valued principles can be.

Commitments of course may falter, and we may run afoul of even the most cherished obligations. The point is that, under normal conditions, expectations of other people's conformity to a moral rule are not a good *reason* to obey it. Nor is it a good reason that others expect me to follow a moral rule. If I find their expectation reasonable, it is because I find the moral norm reasonable; so the reason to obey it must reside in the norm itself. What I am saying goes against the well-known Humean interpretation of our moral obligation to follow the requirements of justice (Hume 1751). This moral obligation is, according to Hume, *conditional* on the expectation that others are following the norms of justice too. In my interpretation, Hume's requirements of justice are social norms, because they fulfill my conditions for a social norm to exist. What distinguishes norms of justice from other social norms is that many of us would have a conditional preference for abiding by such norms because we acknowledge that the normative expectations expressed by Condition 2(b) are *legitimate* and should therefore be satisfied. Their legitimacy may stem from recognizing how important it is for the good functioning of our society to have such norms, but of course their ongoing value depends on widespread conformity. There is nothing inherently good in our fairness norms, above and beyond their role in regulating our ways of allocating and distributing goods and privileges according to the basic structure of our society.[11] However, many of us would feel there is something inherently bad in taking a life, especially when the victim is a close kin. All known societies have developed similar rules against killing one's kin or mating with one's parents. The *unconditional* preference most of us have for not committing such acts may have an evolutionary origin, and typically contemplating killing or incest elicits a strong, negative emotional response of repugnance. What needs to be stressed here is that what makes something a social or a moral norm is our attitude toward it.[12] How we *justify* our conditional or unconditional allegiance has no bearing on the reality of the distinction, and the latter is all that matters to my definition of social norms.

Condition 2 also helps in distinguishing a social norm from a collective habit. People in Pittsburgh wear coats in winter. I expect them to keep

[11] The fact that 'fair' allocations reflect the structure of society is well known to anthropologists. In traditional, authoritarian societies, for example, the allocation of goods is based on rank. Such allocations are accepted by all the involved parties as just (Fiske 1992).

[12] Our attitudes are also shaped in part by the norms that we internalize, which results in a positive feedback loop between attitudes and adherence to norms.

wearing coats in winter and, were anyone interested in my attire, I would say they expect me to wear a coat in winter. But these expectations have no bearing on my decision to wear a coat. There is no connection between my preference for wearing a coat and my expectations about the rest of the population. My not wearing a coat in winter may violate their expectations, and it may cause surprise and puzzlement, but does it matter to my choice? It does not, because I have independent reasons to wear a coat in winter. Condition 2 instead tells that my preference for conformity *depends on* the expectation that others conform, and either the belief that they expect me to conform or the belief that they also prefer me to conform (and may sanction my behavior). Using the language of game theory, we may say that compliance with R is not a strictly dominant strategy.[13] If it were, one would want to follow R irrespective of one's expectation about others' behavior.

Taken together, the conditions I have stated tell us that social norms motivate action, but they do so only indirectly. The direct, underlying motives are the beliefs and desires that support the norm. Thus the presence of a norm of reciprocity, and its salience in a particular situation, motivate me to act in a congruent manner, but my behavior is ultimately explainable only by reference to my preferences and expectations. This statement should not be surprising to those who adopt a methodological individualist perspective. In this perspective, a norm is a social construct reducible to the beliefs and desires of those involved in its practice; if individuals for some reason stopped having those beliefs and desires, the norm would cease to exist.

The conditions for a norm to exist entail, when they are fulfilled, that a social norm is an equilibrium. First, let me briefly define the notion of equilibrium as it is widely used in the social sciences. An *equilibrium* is a situation that involves several individuals or groups, in which each one's action is a best reply to everyone else's action. It is a situation of stable mutual adjustment: Everyone anticipates everyone else's behavior, and all these anticipations turn out to be correct. In other words, an equilibrium is a set of self-fulfilling prophecies that individuals formulate about each other's actions. Social norms, as I stated before, have no reality other than our beliefs that others behave according to them and expect us to

[13] A (strictly) dominant strategy is a strategy that gives the individual who chooses it a better payoff (usually expressed in utility) than any other available strategy. In a game-theoretic context, a (strictly) dominant strategy gives a better payoff than any other available strategy independently of what the other players do.

behave according to them. In equilibrium, such beliefs are confirmed by experience and thus they become more and more ingrained as time goes on. A norm of reciprocity is supported by our beliefs that people will comply with it, and that they expect us to comply with it too. Each time we reciprocate we strengthen the norm and confirm those expectations. In equilibrium everyone reciprocates and is right to do so. But there could also be another equilibrium in which nobody reciprocates. If people expected no reciprocity, there would be no trust in the first place, and again expectations would be self-fulfilling: Everyone would distrust and would be right to do so, because nobody would reciprocate. A situation in which some reciprocate and some do not would not be stable, for the second group might learn that they would do better by reciprocating, and thus switch their strategy (or the first group might learn not to reciprocate, and change their strategy). In some recent work on norm emergence (Bicchieri et al. 2004), I looked at how a norm of trust/reciprocity can emerge in a situation in which different groups display different behaviors, and how they may solidify into an equilibrium. For now, let us agree that social norms, those bundles of self-fulfilling expectations we live by, are equilibria.

If a social norm is followed, then by definition individuals' expectations are self-fulfilling, in the sense that the combination of empirical and normative expectations [Conditions 2(a) and 2(b) or 2(b′)] give one a reason to obey the norm. What sort of reason is this? As I already mentioned, I believe different people may have different reasons for compliance that extend beyond the standard reasons given by many social scientists, namely, that we fear punishment when we disobey a norm. It is certainly possible that some may *fear* the consequences of violating others' normative expectations, because violation may trigger resentment and unpleasant consequences for the transgressor.[14] Such individuals would be motivated to follow a norm to avoid negative sanctions. Yet I would argue that another reason for compliance is the *desire to please* others by doing something others expect and prefer one to do. In this case, the expectation of a positive sanction would be a reason for compliance. A third reason for compliance with a norm is that one accepts others' normative expectations as well founded. In this case, sanctions

[14] This is often the case when members of group A *impose* certain norms on members of group B (the target group). In this case most members of B would conform out of fear of punishment or because of the desire to be rewarded for good behavior. Conditions 2(a) and 2(b′) would in this case refer to expectations about the targeted members of P only.

have no weight. If I recognize your expectations as reasonable, I have a reason to fulfill them. I may still be tempted to do something else contrary to your expectations, but then I would have to justify (if only to myself) my choice by offering alternative good reasons and show how they trump your reasons. This need to offer a justification (to myself as well as others) signals that I recognize others' expectations as cogent. The acceptance of others' expectations as legitimate is usually accompanied by the recognition that negative sanctions against transgressors are also legitimate. If your expectation is reasonable, I must also acknowledge that it is reasonable for you to punish my transgression, even if the reprimand is nothing more than an expression of disapproval of my behavior. The common observation that norm transgression is often accompanied by punishment (or the expectation of punishment) does not entail that norms are only *supported* by sanctions, in the sense that if sanctions were not there, conformity would be entirely absent. Recognizing punishment as legitimate is different from acknowledging that, de facto, violations are punished. The latter does not involve understanding conformity expectations as valid, whereas the former presupposes the acceptance of a norm. It is important to acknowledge that different individuals may need different normative expectations in order to be prepared to obey a norm, and that an individual may follow some norms, but not others, in the absence of any expected sanction.[15]

Fear and the desire to please are powerful motives, but they imply that a norm would only be followed in circumstances in which either there is monitoring of one's actions and sanctioning is possible (as in repeated interaction) or there is some way to ensure that one's action is acknowledged by the people one wants to please or else has a noticeable effect on their well-being.[16] Under anonymity conditions, and when one's action effects are insignificant (as when contributing to some public goods), the motivation to obey a norm would falter. A possible objection to this conclusion is that we may feel *guilt* at violating a norm, and the emotion of guilt supports conformity even in the absence of external monitoring and sanctioning (Elster 1989). According to this view, emotions directly *cause* conformity. But why and when do we feel guilt? Imagine a situation in which someone does not expect others to conform to a practice of truth-telling. He has observed people openly lying and has been lied

[15] In Chapter 3, I discuss the differences we observe in the behavior of Proposers in Ultimatum versus Dictator games.

[16] Individuals differ as to the scope of people they want to please. Most of us stop at family and friends, but some may include acquaintances and even strangers.

to often enough to expect further dishonesty. Yet he is made to believe that he is expected to conform to a norm of truth-telling. It is likely that this individual would consider the expectation illegitimate, and he would feel no guilt at violating it. Guilt, as well as resentment, presuppose the violation of expectations we consider *legitimate*. It is irrational to resent a malfunctioning computer, but it is reasonable to resent the seller if we think he should have known (and told us) the computer was defective. We trusted him, and he flouted our legitimate expectations of honesty and good faith. Guilt and resentment *signal* that a social norm is in place and that mutual conformity expectations are legitimate. It is reasonable to feel guilt or resentment precisely because there is a norm, a set of mutual expectations that we recognize should be met. The existence of an accepted norm that one contemplates violating is the source of guilt, but it is the recognized legitimacy of mutual normative expectations, not the emotion of guilt, that motivates conformity.

Notice that I am not postulating a generic desire to meet, whenever possible, other people's expectations. In his analysis of conventions, Sugden (2000) assumed we possess a 'natural aversion' toward acting contrary to the preferences (and expectations) of others. This propensity may be true for the preferences and expectations of family and friends, but it is hardly at work with strangers. As social distance increases, we tend to care less and less for others' preferences and expectations, especially when these preferences and expectations run counter to other interests we have. Sugden's assumption would restrict norm-abiding behavior to a circle of family and friends, but these are precisely the circumstances where norms are not needed. In large, anonymous groups, if we do not want to act contrary to others' normative expectations, it must be because we find such expectations reasonable. The acceptance of others' normative expectations as reasonable is the third kind of motive to conform to a social norm I mentioned before. This need not be a motive for everyone, but in all cases in which anonymity and the absence of sanctions tempt us to defect, for a norm to survive there must be a critical number of people for which such reasons have power.

Since social norms often go against our self-interest, especially if we narrowly interpret it as a desire for material possessions, a social norm need not be an equilibrium of an ordinary game in which payoffs represent self-interested preferences. Thus, for example, a cooperative norm cannot be a Nash equilibrium of the PD game represented in Figure 1.1.[17] If

[17] A Nash equilibrium is a combination of strategies, one for each player, such that each player's strategy is a best reply to the strategies played by the other players.

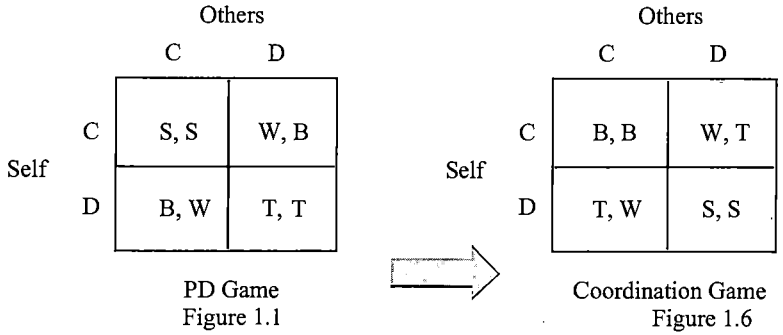


FIGURE 1.6. Norms transform games

such a norm exists and is followed, however, the original PD game would be transformed (at least for the norm followers) into the subsequent, very different game shown in Figure 1.6.

In the traditional Prisoner's Dilemma game, each player's preference ranking is DC > CC > DD > CD. As before, B stands for 'best,' S for 'second best,' and so on. In the symmetric game of Figure 1.6 instead, each norm follower's preference ranking is CC > DD > DC > CD. That is, the players who follow a cooperative norm will do it because their empirical and normative expectations have been met and hence they *prefer* to obey the norm. The new game in Figure 1.6 is a coordination game with two *strict* Nash equilibria, one of which is Pareto superior to the other.[18,19] When a norm of cooperation is obeyed, a game like the PD of Figure 1.1 is *transformed into a coordination game*: Players' payoffs in the new game will differ from the payoffs of the original game, because their preferences and beliefs will be as in Conditions 2, 2(a), and 2(b) or 2(b′) previously outlined. Indeed, if a player knows that a cooperative norm exists and expects a sizeable part of the population to follow it, then, provided she also believes she is expected (and maybe also preferred) to follow such norm, she will have a preference to conform to the norm in a situation in which she has the choice to cooperate or to defect. Note that what I am saying implies that a social norm, unlike a convention, is never a solution

[18] In a strict Nash equilibrium each player's strategy is a unique best reply to the other players' strategies. This means that a strict Nash equilibrium cannot include weakly dominated strategies.

[19] A coordination game is a game in which there are at least two Nash equilibria in pure strategies, and players have a mutual interest in reaching one of these equilibria (CC or DD in Figure 1.6), even if different players may prefer different equilibria (which is not the case in Figure 1.6).

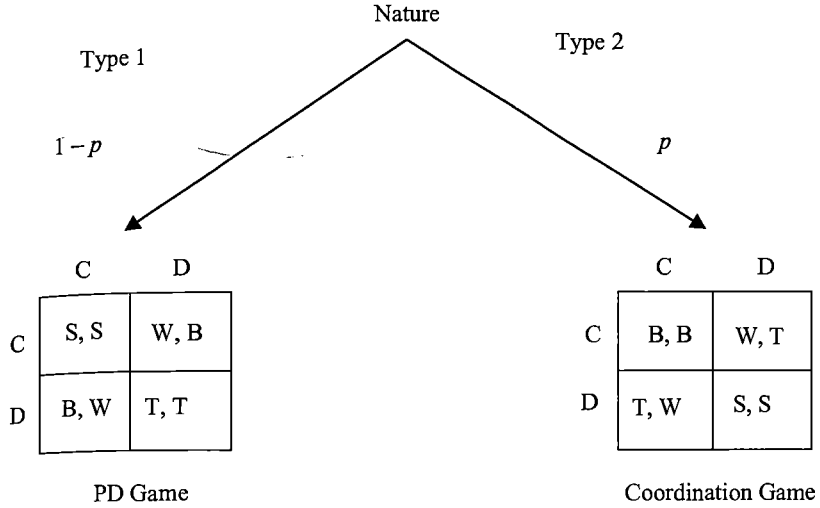


FIGURE 1.7. A Bayesian game

of an original coordination game, though it is an equilibrium of the new, transformed game it *creates*.

It is important to recall that my definition of social norm does not entail that *everybody* conforms. In fact, the definition says that a social norm may exist and not be followed. For some, the PD game of Figure 1.1 is never transformed into any other game. And even a person who starts playing a coordination game like the one in Figure 1.6 may revert to playing the regular PD game if she realizes that Condition 2(a) (*empirical expectations*) is violated. Let me clarify this point with a simple example. Suppose an actor is faced with a finitely repeated PD, and suppose the situation is such that a 'cooperative' norm is primed.[20] The player knows there exists a cooperative norm that applies to this kind of situation. The player also knows that there are several types of players, some of which would not see the game as he does. To make matters easy, suppose there are two types of players, those who simply see the game as a PD and those who follow a cooperative norm.[21] In this case we may model the choice situation as a Bayesian game (Figure 1.7) in which Nature picks a player type with a given probability, so that with prior probability p the opponent one faces

[20] In Chapter 4 I discuss in detail how such 'cooperative' norms might be primed.

[21] In a finitely repeated game, even a 'selfish' player may want to cooperate for a while, if it is not common knowledge that all players are rational and selfish (Kreps et al. 1982). This consideration, however, has no bearing on my argument because, until a defection is observed, a player cannot distinguish between a forward-thinking selfish type and a true cooperator.

is playing a coordination game, and with probability $(1 - p)$ he is playing a PD game.[22] If a norm-follower assesses a sufficiently high probability to being matched with a similar type, he will cooperate.[23]

When faced with a defection, however, the player will reassess his probabilities and possibly revert to playing the equilibrium strategy (defect) for the traditional PD game. One might thus say that the *existence* of a norm always presents a conditional follower with a Bayesian game: If the normative and empirical expectations conditions are fulfilled, she will assess a higher probability to being matched with a similar player type (a norm follower) and act accordingly. But she must also be prepared to revise her probabilistic assessment in case experience contravenes her previous expectations.[24] Note that the existence of a social norm facilitates equilibrium selection in the Bayesian game faced by the conditional norm followers. If the probability of being matched with a similar type is high enough, C,C is the selected equilibrium; otherwise D,D will be selected. (Appendix 1 presents a formal treatment of a norm-based utility function and the conditions under which a PD game becomes a coordination game.)

This simple and elegant game-theoretic model offers a language, built on the notions of belief and preference, in which to cast what we commonly observe: In an experimental setting in which repetitions of a PD-like game are allowed, we witness high initial levels of cooperation. Yet cooperation precipitously declines as soon as some players defect (Fehr and Gachter 2000a). Whether a game-theoretic model provides an acceptable explanation for what we observe depends in part on our willingness to take 'as if' models seriously, which in turn relates to the possibility of drawing interesting predictions from them. In the case at hand, people may not be aware of their preferences and never have made a probabilistic assessment of the situation; yet, if we take their behavior to reveal certain dispositions, we may predict that, *ceteris paribus*, factors that we expect will change their expectations will have a measurable effect on future choices.

[22] When players are uncertain as to the type of player they are facing, they will assess some probability that the other player is of a certain type. Typically the list of all possible types and their prior probability of occurring in the population are taken to be common knowledge among the players (Harsanyi 1967, 1968).

[23] If players use an availability heuristic to come to this probability assessment, the probability of playing a coordination game might initially be much higher. That is, if a player is the type who follows a cooperative norm, he tends to believe there is a high probability that others are like him.

[24] This revision is governed by a "learning rule." I discuss one such rule in Chapter 6.

Descriptive Norms

Let us now look at how the definition of social norms given above differentiates several types of social constructs and behaviors that are often lumped together. Sometimes 'norm' means what people commonly do in certain situations, what constitutes 'normal' or 'regular' behavior. This notion of regular behavior differs in important respects both from a shared habit and from what people believe ought to be done, what is socially approved or disapproved. The regular behaviors I am referring to, and their influence on people's choices, have been extensively studied by social psychologists, most notably Cialdini et al. (1990), who dubbed them *descriptive* norms. Examples of descriptive norms are all sorts of fashions and fads, in addition to the many collective behaviors that people (rightly or wrongly) deem to convey important information about the surrounding world. Conventions, as we shall see, are a kind of descriptive norm, but not all descriptive norms become stable conventions. Note that there is no intrinsic property of a behavioral pattern that makes it a descriptive norm: What is a descriptive norm for one group may be an entrenched social norm for another. Dress codes are a case in point. For the office workers at a particular firm, a 'dress-down Friday' informal rule is nothing more than a fashion code that, though widely adopted, remains entirely discretionary. For teenage members of a Los Angeles gang, on the contrary, a dress code may signal group loyalty, so much so that every member is expected to rigidly adhere to the code and transgressions are punished. What makes a collective behavior a descriptive or a social norm are the expectations and motives of the people involved. This point is worth emphasizing: It is the way we relate to behavioral rules by way of preferences and expectations that gives them their identity as habits, norms, or mere conventions.

We conform to social norms because we have *reasons* to fulfill others' normative expectations. These reasons often conflict with our self-interest, at least narrowly defined. Conformity to descriptive norms is, on the contrary, *always* dictated by self-interest: We conform because such norms make life easier for us, because we want to 'fit in' or do the right thing – as when we adopt a new fashion – or simply because they provide evidence of what is likely to be effective, adaptive behavior, as when we bought Internet stocks because many people we know were buying them and were doing well. Often there are good prudential or informational reasons to "do as the Romans do." Conformity to a descriptive norm may be motivated by a desire to imitate others' behavior in uncertain or

ambiguous situations. In such circumstances, others' behavior provides us with information about the appropriate course of action, as when a young employee imitates older, more experienced colleagues' way of handling complaints. Imitation may be a reasonable, cost-effective choice, provided we believe that the majority's behavior or opinion conveys the information we lack. There are many occasions in which we have to make a quick decision without much information about the environment: Gathering information may be unfeasible or have too high an opportunity cost in terms of resources (such as time and money) that one would more effectively employ elsewhere. Or we may be in a condition in which the wrong decision could have serious consequences, and we lack the expertise to properly evaluate the situation. Conversely, there are circumstances in which the consequences of a decision are not too important, and here again gathering information seems a waste of resources. In all these cases, we look at the choices other people make as a guide to our own choices. While this may seem like a good deal for the actors at the time, it can ultimately mean all actors depend on the choices of one (or a few) first mover(s), and those choices may or may not be good ones. This type of 'informational cascade' (Banerjee 1992) may be the reason why some inefficient descriptive norms emerge and persist, as I will discuss in Chapter 5.

For now let me stress that conformity to a descriptive norm need not involve an obligation or normative expectations: We do not feel any group pressure to conform, nor do we believe that others expect us to comply with what appears to be a collective behavior. Deviation from the 'norm' is not punished, nor is compliance overtly approved. For example, if I decide – alone among my friends and co-workers – not to invest my retirement money in stocks, I do not expect to be blamed or ostracized. At worst, they will think I am overly cautious. A crucial feature of *descriptive norms* is thus that they entail *unilateral* expectations. Though we may have come to expect others to follow a regular behavioral pattern, we do not feel any social pressure to conform. That is, Conditions 1, 2, and 2(a) apply but Conditions 2(b) and 2(b′) do not. In most cases of descriptive norms, there simply are no reciprocal expectations: We do not believe others care about our choices or expect us to follow any particular behavior. When I choose to adopt a new fashion, I usually do not think I am *expected* to follow it. But even in those cases in which we are aware that we might be expected to follow the majority's decision or opinion, we do not count on being blamed if we follow a different path. Others may think it would be prudent or reasonable for us to behave as they do (for example,

to pick a certain stock portfolio), but the 'ought' involved in stating prudential reasons is very different from a normative ought. I might recognize the reasonableness of others' expectations, but not their legitimacy.

Fulfilling others' expectations in this case is not a reason for compliance, whereas expecting a majority of people to behave in a given way is a *necessary reason* to adopt that behavior. It is only a necessary reason, however, because one must also have a *conditional preference* for conforming. Expectations alone cannot motivate a choice: My choice to conform depends on expecting a majority of people to conform, but it must be that I prefer to follow such 'normal' behavior on condition that it is the majority's behavior. This latter condition differentiates a descriptive norm from a collective habit: In the example of wearing a coat in the Pittsburgh winter, I have an independent reason (and thus a preference) to wear a coat, irrespective of what other people do. But in the case of a new fashion, following it depends on one's perception of what other people do. After Mary Quant introduced the miniskirt in the 1960s, it probably took a small number of trendsetters to reach a critical mass and start what became a major change in women's fashion. That critical mass of women, however, was crucial in determining the success of the new attire: Most women would not have started wearing a miniskirt if not for the sense that it was now 'in' and many celebrities were wearing it. It should be noted that often it is the *perception* of a critical mass, rather than a real critical mass, that tips the balance in favor of the new behavior. A small but vocal minority, or an endorsement by some celebrity, may thus be enough to induce a change in mass behavior.

The conditional preference for conformity may be dictated, among other things, by a desire to 'fit in' or be fashionable, or just by prudential reasons; it does not, however, spring from a desire to fulfill other people's expectations or from fear of being punished if one does not meet them. For a descriptive norm to exist, the following conditions must be met.

Conditions for a Descriptive Norm to Exist

Let R be a *behavioral rule* for situations of type S, where S is a coordination game. We say that R is a descriptive norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$,

1. *Contingency*: i knows that a rule R exists and applies to situations of type S;
2. *Conditional preference*: i prefers to conform to R in situations of type S on the condition that:

(a) *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S.

A descriptive norm is followed by population P if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for all $i \in P_f$, Condition 2(a) is met for i and as a result i prefers to conform to R in situations of type S.

A descriptive norm thus tells what a person would do if he had certain expectations. For instance, "walk on the left side of the sidewalk" and "walk on the right side of the sidewalk" are both descriptive norms. Some people may follow the first rule (because they expect others to do the same), some people may follow the second rule (again, because they expect others to do the same), and some people may follow neither rule (because they do not expect a sufficient number of other people to walk on a specific side of the sidewalk). Even in a society where one of the rules has been conventionalized, it is clear that the other rule still exists as a possibility: I drive on the right side of the road, but if I observed large numbers of people driving on the left side of a particular road, my expectations would change and I would consider driving on the left side of that road.

As in the case of social norms, the preference for conformity is conditional, but this time it is only conditional on expecting others to follow the behavioral rule in a given class of situations. Note that a descriptive norm that is followed is an equilibrium, in the sense that followers' beliefs will be self-fulfilling: If one believes R to be widely followed, then it is in one's interest to follow R, too. Thus, if enough people come to believe R is the 'norm,' they will behave in ways that further validate those beliefs. The conditional preference for conformity may be driven by the desire to imitate those we believe are more informed or by the hope of 'fitting in' a group we value. Or it may simply be the wish of doing what we think most people do. Be it as it may, a preference for conformity depends on expecting others to conform to R.

If a descriptive norm is an equilibrium, what sort of game is it an equilibrium of? Consider again the miniskirt fashion: In this case, a woman has several choices of attire, of which the miniskirt is one. Assume for simplicity that there are only three possible types of clothes women can choose from: M (miniskirt), L (long skirt), and P (pants). Assume also that a woman already has L and P in her wardrobe, and has to decide whether to buy and wear M. The choice of M is thus more costly than L or P, but she prefers above all to be fashionable. Her choice matrix would look like the one in Figure 1.8.

		Others: M	Others: L	Others: P
Self	M	1, 2	0, 1	0, 1
	L	0, 2	2, 1	0, 1
	P	0, 2	0, 1	2, 1

FIGURE 1.8. An Imitation game

Notice that the payoffs of 'Others' need not be the same as the payoffs of 'Self.' Indeed, suppose 'Others' are the trendsetters that start a new fashion. I assume the trendsetters will not care whether 'Self' follows the new fashion; what the trendsetters care about is self-expression, and starting a new fashion is not their goal [M may have a higher payoff (2) for trendsetters because they always prefer to do the 'new' thing]. 'Self' instead wants to imitate the trendsetters; hence she cares about whether she coordinates with 'Others.' Because 'Others' may not care at all about being imitated, the three Nash equilibria of the game are not strict. Imitation is a one-sided coordination game.[25] Even if the choice of M is more costly than P or L for 'Self' (it has a lower payoff), if it is believed that now "it is *in* wearing miniskirts," the imitators' choices will converge to M. The example shows that a descriptive norm may be a suboptimal equilibrium and still be the one chosen by the players. It also shows that the class of games of which descriptive norms are equilibria is much larger than the class of coordination games of which conventions are equilibria. As I shall discuss shortly, the latter are always coordination games without nonstrict Nash equilibria and for that reason, in such games, all players prefer that everybody conforms. Such preferences are absent in a descriptive norm.

Earlier I represented social norms as coordination games, too. There is a crucial difference, though. The existence of a social norm *transforms* a game like the Prisoner's Dilemma (or any other mixed-motive game) into

[25] Note that 'most other women' need not refer to an entire population or even a large group. Some descriptive norms are *exclusive*, in that they signal belonging to a special, selected group. Fashion may play that role on occasion.

a coordination game (or a Bayesian game, in which we may be playing a coordination game with a given probability), by providing actors with an alternative set of expectations and preferences. But the problem that a social norm is solving in the first place is *never* a coordination problem. If I expect everybody to cooperate, to be fair, or to reciprocate favors, I may be tempted not to, and only a desire to fulfill others' expectations may induce me not to stray. This desire may spring from fear, benevolence, or the acknowledgment of the legitimacy of others' reasons and expectations. Social norms by and large apply to situations in which there is a conflict between selfish and pro-social incentives. In contrast, descriptive norms *solve* a preexisting coordination problem (even if it is a unilateral one, as in imitation). If so, following a descriptive norm is not in opposition to self-interest. Indeed, it is usually in one's self-interest (however narrowly defined) to follow a descriptive norm. In sum, we may say that a descriptive norm is always an equilibrium strategy of an original coordination game. In a given situation S, a descriptive norm is followed if and only if the players of the coordination game expect (with sufficiently high probability) a particular equilibrium strategy to be played, and thus they play that strategy as well.[26]

Note that the game-theoretic representation is silent about the dynamics leading to one particular equilibrium. We still need a plausible story about the dynamics that led women to adopt *en masse* the miniskirt. For the moment, however, this should not be a matter of concern; for now all we want to answer are questions about conformity and norm elicitation.

Conventions

Descriptive norms, such as fashions and fads, can wane rather quickly, but some of them may crystallize into stable *conventions*, such as signaling systems or dressing codes. Such conventions are useful because they coordinate our expectations and often act as signals that facilitate interaction and communication. Usually no intrinsic value is attributed to a convention, although violating it can be costly, as the cost is directly related to the consequences of breaching a coordination mechanism. For example, the trader on the stock exchange floor signals with her fingers how many shares she wants to buy or sell. The failure to do so is not socially

[26] A definition of descriptive norms that *requires* them to be followed would limit descriptive norms to a time-varying and imprecisely defined subset of the equilibria and would also make it hard to talk about equilibrium strategies that are not currently being played.

condemned, sanctioned, or accompanied by guilt. Not following the convention simply means the trader will not be able to communicate what she wants and lose an opportunity to gain. When a convention is in place, expectations of compliance are *mutual*. An actor expects others to follow the convention, and she also believes she is expected to follow it by the other participants in the conventional practice. The traders expect each other to follow the signaling convention, as much as we normally expect a competent speaker to stick to the rules of English usage. Yet such mutual expectations are never a sufficient reason to adhere to a convention. It must be that one has a conditional preference for coordinating and communicating with others, as failure to coordinate and communicate comes with a personal cost.

David Lewis first defined conventions as equilibria of coordination games (Lewis 1969). According to Lewis, a convention is a regular pattern of behavior that is a strict Nash equilibrium in a coordination game with $n \geq 2$ strict Nash equilibria.[27] This requirement is meant to capture the *arbitrariness* of conventions, in particular the awareness on the part of those participating in a convention that there are possible alternative arrangements. In a coordination game, the interests of the participants may or may not perfectly coincide. In the miniskirt example, all the followers had the same (ordinal) preferences. In the game in Figure 1.9, instead, the players' interests do not exactly coincide. What matters though is that everyone does better by coordinating with the choices of other players than by 'going solo.'

The game in Figure 1.9 can be interpreted as a situation in which two people want to coordinate or 'be together,' but one would prefer to go to the Opera whereas the other prefers playing Golf. The game has two strict Nash equilibria in pure strategies, (Golf, Golf) and (Opera, Opera), and a mixed-strategy equilibrium in which 'Other' chooses Golf with probability 1/3 and Opera with probability 2/3, and 'Self' chooses

[27] Lewis's account of convention is quite different from mine, and it runs as follows (p. 78):

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P, (1) almost everyone conforms to R; (2) almost everyone expects almost everyone else to conform to R; (3) almost everyone has approximately the same preferences regarding all possible combinations of actions; (4) almost everyone prefers that any one more conform to R, on condition that almost everyone conforms to R; (5) almost everyone would prefer that any one more conform to R', on condition that almost everyone conform to R', where R' is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R' and to R.

		Other: Golf	Other: Opera
Self	Golf	1, 2	0, 0
	Opera	0, 0	2, 1

FIGURE 1.9. A Coordination game

Golf with probability 2/3 and Opera with probability 1/3. Clearly the preferences of the players are not identical. They do, however, prefer to be together rather than be separate. The players may settle on one of the equilibria for whatever reason, but once they are in equilibrium, they have no incentive to deviate from it.[28] When I say that a convention is 'self-sustaining,' I just mean that each actor has a self-interested motivation to conform to the convention.

The matrix in Figure 1.9 does not tell us *which* equilibrium is played, because it all depends on the expectations players bring to the game. Thus 'Self' may have to settle for Golf, if he expects 'Other' to make that choice, and vice versa. But how are these expectations justified? This is a well-known problem in game theory: Even if players have common knowledge of the structure of the game and of their mutual rationality, usually this information is not sufficient to select a particular equilibrium strategy (Bicchieri 1993). In this case, we must introduce some salience criterion of choice, and common knowledge thereof, to solve the equilibrium selection problem. Salience may be provided by precedent or by an explicit agreement. Lewis (1969) unambiguously referred to precedent as a mechanism by which players succeed in coordinating on one particular equilibrium. Schelling (1960), on the other hand, referred to focal points. However, salience and focal points are not satisfactory solutions, because for them to do their coordination job it must be common knowledge among the players that they describe the game in the same way; but unless it is explicitly assumed, there is no reason to believe that common knowledge exists. This interpretation of the coordination

[28] An account of how a convention emerges would look at repetitions of the stage game depicted in Figure 1.9. The convention in this case might be that people alternate between Golf and Opera, or that they do one or the other with fixed probabilities.

game in Figure 1.9 is a static, stylized description of the conditions under which a convention is likely to *emerge*, not an analysis of how players attain common knowledge of the shared criteria that will help them solve their coordination problem.

Another possible interpretation of the game in Figure 1.9 is that one of the two equilibria has already been selected and consequently a convention is in place. 'Self' will know, for example, that in situations of type *S* almost everyone chooses to play golf. She thus has an empirical expectation about what 'Other' will do and a conditional preference for conformity given her expectation. In this case 'Self' will conform and, if 'Other' has a similar expectation and preference, he will follow the established convention, too. In this case no common knowledge is necessary for players to play the 'play golf' equilibrium: First-order expectations are all that is needed. This interpretation of the game refers to the *survival* of a convention: A convention persists if agents have the right kind of empirical expectations. The question now becomes how agents come to form such expectations, or reason inductively from past cases. For example, when 'Self' is faced with situation **s**, she will look for analogies with past situations she has experienced and eventually decide there are enough relevant similarities to categorize **s** as a member of *S*, the class of situations to which a given behavioral regularity applies. The next step for 'Self' is to decide that that particular behavioral regularity can be projected as a genuine regularity; otherwise she would have no reason to expect it to persist. Sugden and Cubitt (2003) point out how Lewis explicitly recognized that inductive inferences are crucial in maintaining a convention and offer a formal model of Lewis's informal description of how common knowledge that a behavioral regularity will persist is attained. Without entering into the details of Sugden and Cubitt's formal reconstruction of Lewis's argument, let me point out that Lewis's argument is crucially dependent on assuming shared inductive standards, and one of these standards is the common recognition that only certain behavioral patterns can be projected. In the next chapter I address the problem of what grounds inductive inferences (especially inferences about social behavior); for now let me point out that a game-theoretic account of norms and conventions, insofar as it describes them as equilibria of particular types of games, is both inescapably static and epistemically inadequate. Not only do we need dynamic accounts of how norms and conventions emerge, but also a better understanding of the kinds of cognitive capabilities that allow us to recognize and project behavioral patterns as such. I will address both issues in later chapters.

All we need to emphasize for the moment is that a convention is a realized equilibrium of an *original* coordination game without nonstrict Nash equilibria, and that it is in a player's self-interest to stick to it. We are now ready to give a more precise definition of convention that hopefully captures its characteristic features.

Conditions for Conventions to Exist

A descriptive norm is a convention if there exists a sufficiently large subset $P_f \subseteq P$ such that, for each individual $i \in P_f$, the following conditions hold:

1. *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S and
2. S is a coordination game without nonstrict Nash equilibria.

Recall that, for a descriptive norm to be followed, empirical expectations [Condition 2(a)] had to be met. Hence, a convention is always a *followed* descriptive norm, because empirical expectations are met. That is, the follower of a convention always expects a sufficiently large subset of P to conform. Note that a descriptive norm could be a nonstrict Nash equilibrium; a follower could imitate a trendsetter, but the latter would not be interested in coordinating with the 'followers.' In the case of a convention, instead, there is no such indifference.

There are several important differences between conventions and social norms. One is that conventions, in order to exist, have to be followed. Social norms (and descriptive norms) instead can exist without being followed. Second, one conforms to a convention because of the belief that others behave in the expected way, because it makes sense to follow a convention only if there is reasonable certainty that it is still in place. Conforming to a social norm, on the contrary, requires that *both* normative and empirical expectations are met. Because conventions do not run counter to selfish motives, but social norms often do, if only empirical expectations were fulfilled, one would have a reason to follow a convention, but he would be seriously tempted *not* to conform to a social norm. In both cases, the players are playing a coordination game without nonstrict Nash equilibria, but whereas a convention solves an original coordination game, a norm transforms (with a certain probability) an original mixed-motive game into a coordination game and at the same time helps players to select one equilibrium.

The neat boundaries I drew between descriptive norms, conventions, and social norms are quite blurred in real life: Often what is a convention

to some is a social norm to others, and what starts as a descriptive norm may in time become a stable social norm. Sometimes (but by no means always) the passage is marked by the presence of a new preference for universal conformity. In the trading example, the trader does not prefer that every other trader follow that specific signaling convention. Of course, she prefers that there is a signaling system, but she does not care if some traders do not follow it (provided the system is still in place). If another trader suddenly decides to make different signals, he is the only one to bear the cost of deviating from the conventional sign language. The case of traffic rules, the quintessential example of a convention, is quite different. Driving according to 'personal' rules may cause severe damage. The reckless driver is prone to cause accidents involving other people, who thus have to bear the costs of his infraction. When breaking a convention creates negative externalities, people prefer not just that the convention is in place, but also that everyone follows it. Such violations are usually legally sanctioned, but, even more importantly, they are also informally sanctioned by society. A reckless driver is blamed as irresponsible: We think he should have observed traffic rules. When breaking a coordination mechanism produces negative externalities, we may expect conventions to become full social norms.

A good example of such a transformation would be the stag-hunt game (Hume 1739). In this game, the hunters could coordinate their efforts and get a stag, which is a much bigger and valuable prey than a hare, which they could hunt alone and get with certainty. The game can be represented as shown in Figure 1.10.

If the players agree to hunt the stag together, they may get a better payoff (2) than hunting alone (1). However, even if a stag-hunting convention is in place, the larger the number of hunters, the higher the

		Others: Stag	Others: Hare
Self	Stag	2, 2	0, 1
	Hare	1, 0	1, 1

FIGURE 1.10. The stag-hunt game

probability that someone might deviate from it. The (Stag, Stag) equilibrium, though Pareto dominant, is risky because, if someone deviates from it, Self risks remaining empty-handed. The (Hare, Hare) equilibrium is risk-dominant, because by hunting hares alone success is guaranteed.[29] Thus, if p (Others play S) is greater or equal to 1/2, Self will choose Stag; otherwise she will choose Hare. In this case the players might agree to impose sanctions on the lone hunters, especially when the hunting group is small and even a single deviation risks preventing the stag from being successfully hunted. What started as a convention may thus in time become a full social norm.

This does not mean that a social norm is in place *because* it prevents negative externalities from occurring. Many social norms are not the outcome of a plan or a conscious decision to enact them; they emerge by human action but not by human design. Some conventions may not involve externalities, at least initially, but they may become so well entrenched that people start attaching value to them. For example, a group of people may routinely avoid smoking before there arises a consensus disapproving this behavior. Once a public consensus is reached, smoking incurs new costs. Not only would one be expected not to smoke, but the occasional smoker would incur the blame of the entire group. At this point, a social norm is born. It may also happen that some conventions lend themselves to purposes they did not have when they were established. Norbert Elias (1978) illustrated how rules of etiquette, such as proper ways to eat and drink, developed to become a sign of aristocratic upbringing and refinement, and were effectively used to exclude those who did not belong to the ruling class. Thus a thirteenth-century peasant, and even a city burgher, would be excused if he slurped with his spoon when in company, drank from the dish, or gnawed a bone and then put it back on the communal dish. It would have come as no surprise if the ill-bred blew his nose in the tablecloth, poked his teeth with the knife, and slobbered while he drank, but no nobleman was allowed such lack of manners. Definitions of socially unacceptable behavior, or 'coarse manners,' were uniformly shared by thirteenth-century writings on table manners – simultaneously appearing in Italy, Germany, and England – that recorded for the first time a long-standing oral tradition reflecting what was customary in society. The standard of good behavior promoted in these works is the behavior of the aristocracy, the courtly circles gathering around the great feudal lords. Social differences were much more important than they are today, and

[29] For a definition of risk-dominance, see Harsanyi and Selten (1988).

they were given unambiguous expression in social conduct. Because at that time eating together was a significant moment of socialization, table manners came to play an essential role in shaping the identity of the aristocracy. A member of the ruling class was *identified* as such through his 'courtesy' or good manners. Had he not respected the rules of etiquette, he would have been met with contempt and perceived as threatening the established class boundaries.

Another example of a convention that evolved to become an important social signaling device is footbinding in China (Mackie 1996). The practice of footbinding might have been invented by a dancer in the palace of the Southern T'ang emperor, or it may have originated among slave traders as a restraint on female slaves, but it soon spread to all but the lowest classes in the population, becoming a sign of gentility and modesty and an essential condition for marriage. A family that did not impose such painful mutilation on its female children would have come to signal, among other things, a dangerous disregard for tradition and custom. As a consequence, it would have been ostracized and its young females regarded as unsuitable mates. Given enough time, what starts as a descriptive norm may become a stable convention. And conventions that prevent negative externalities, or those that come to fulfill an important signaling function, especially when the signal is related to social status or power, are easily amenable to being transformed into social norms.

There are many rules of social interaction we usually think of as mere conventions but, on closer inspection, show all the characteristics of social norms. These rules have become so entrenched in the texture of our lives, so imbued with social meanings, that we cannot ignore them with impunity. Everyday life is rife with implicit conventions directing the way we speak, walk, make eye contact, and keep a distance from other people. We are seldom aware of them until they are broken; however, when they are breached we may experience anger, outrage, and confusion. A person who speaks too loudly, stands too close, or touches us in unexpected ways is usually perceived as disturbing and offensive, if not outright frightening. Cultures differ in setting the boundaries of personal space, but once these boundaries are in place, they define 'normal' interactions, help in predicting others' behavior, and assign meaning to it. The rules that shape the perimeter of our personal sphere thus have an important signaling function: We resent those who trespass these boundaries precisely because we perceive those individuals as being hostile and threatening. Conventions of public decorum, such as manners and etiquette, are more explicit but not less important because, among other things, they signal

respect for others and for social relationships. Breaching them can offend and bring forth retaliation. Simmel's example of the dangers of failing to greet an acquaintance on the street underscores this point: "Greeting someone on the street proves no esteem whatever, but failure to do so conclusively proves the opposite. The forms of courtesy fail as symbols of positive, inner attitudes, but they are most useful in documenting negative ones, since even the slightest omission can radically and definitely alter our relation to a person."[30] When a conventional manner of interaction has acquired such an important social meaning, we would rather refer to it as a social norm. Such norms, as opposed to conventions, are accompanied by what are perceived as legitimate expectations of compliance: We feel almost entitled to a courteous greeting, and the annoyance and resentment we direct against those who willingly ignore us indicate we are in the realm of normative expectations.

Following Social Norms

Social norms prescribe or proscribe behavior; they entail obligations and are supported by normative expectations. Not only do we expect others to conform to a social norm; we are also aware that we are expected to conform, and *both* these expectations are necessary reasons to comply with the norm. Contrary to what happens with descriptive norms and conventions, being expected (and preferred) to conform to a social norm may also give us a sufficient reason to conform. I have mentioned fear, benevolence, and the desire to fulfill others' legitimate expectations as three different reasons why normative expectations (and preferences) matter to conformity. Fear should never be discounted, because there are many cases in which one obeys a norm only because neglecting others' expectations and preferences will bring about some form of punishment. We may conform without attributing any intrinsic value to the norm and without finding others' expectations legitimate. Some Arab women may observe Muslim sexual mores, and Corsican men embrace norms of revenge, for fear of being punished if they break the rules. In both cases, they may find their community norms oppressive and ill-suited to modern life, but whoever speaks or rebels first runs the risk of bearing huge costs. Breaking the rules looks like the risky cooperative choice in a social dilemma. Freedom from a bad norm is a public good that is often very difficult to bring about.

[30] Cf. G. Simmel (1950, p. 400).

At the opposite end of the spectrum are those who conform because they attribute some value to what the norm stands for. People vary in the degree to which they are prepared to stand for a given norm. Some of us value a rule of reciprocity, because we see how it helps society function smoothly, but we would be prepared to shed the rule in an environment where it is consistently violated. Others might find deep moral reasons for upholding it even in the face of betrayals. A thirteenth-century member of the ruling class would have refrained from blowing his nose in the tablecloth because that behavior was not 'courtly' or appropriate for a nobleman. Nowadays most of us would be ashamed at displaying such bad manners in front of a table companion. Even if alone, we tend to avoid this kind of behavior, finding it not just unsanitary but also a little demeaning. The negative social sanctions that may follow a transgression are usually reasons for compliance when a social norm is not well established. But later, when the norm has become a well-entrenched practice and we have come to attribute a certain virtue to what it prescribes, external sanctions seldom play a role in inducing conformity. Thus a smoker who avoids smoking in public places for fear of being reprimanded may in time come to see the merit of this policy and refrain from smoking in public places even when alone. Philosophers have pointed out that it is a fallacy to infer *ought* from *is*, but personal as well as historical evidence tells us that we are readily victims of this 'naturalistic fallacy': When a practice is well entrenched, we often come to attribute to it some intrinsic value. In such cases we recognize the legitimacy of others' expectations and feel an obligation to fulfill them.

Neither the person who obeys a norm because a reward or a punishment is in place, nor the person who always obeys out of a deep conviction of the norm's merits presents us with a particular problem. Sometimes, however, we follow social norms even in the absence of external sanctions: Our choices are anonymous, and we are reasonably sure nobody is going to monitor us and detect behavior that runs counter to the norm. Even if a choice is not strictly anonymous, there are many cases in which we can easily turn our backs to the situation and leave without risking any penalty. When we leave a tip at the diner sitting along the motorway we happen to be passing, we are behaving like regular customers even if this is the first and probably the last time we will see that waiter, so there is no obvious punishment or reward in place. It might be argued that in this case we are in the grip of personal norms and would experience guilt or shame were we to transgress our self-imposed rules. If this were the case, we should observe consistent compliance with a tipping norm in a variety

of circumstances, but often the same individual who is ready to leave a tip at the diner may not do so when in a foreign country, even in those cases in which the 'service included' clause is not present. The same inconstancy we encounter with tipping may occur with respect to much more important social norms, such as those regulating fair division or reciprocation. People who reciprocate on one occasion may avoid reciprocating on others without apparent reason. I am not referring here to cases in which it is acceptable to transgress a norm.[31] For almost every norm one can think of, there are socially acceptable exceptions to it. Thus I am normally expected to return favors, but an intervening hardship may excuse me; similarly, many would deem it inappropriate to return a favor that was not requested and looks like a veiled bribe. The cases of interest are rather those in which one is expected to adhere to a norm and does not, but we have evidence that on other, similar occasions, the same person complied with the norm even in the absence of any obvious sanction. I am interested in explaining such apparent inconsistencies across and within individuals.

The brief taxonomy of norms I have proposed is of some help here, because what is baffling is not inconsistency in following a descriptive norm or a convention, but the inconsistency we experience with regard to social norms. For example, when a coordinating convention is in place, it is in everybody's interest to follow it, and when we observe inconsistent behavior we are likely to attribute it either to a misunderstanding of the situation or to poor learning about how and when to follow the convention. Whenever I go back to England, I have to pay special attention the first few days I drive a car, because driving on the left side of the road feels unnatural. If I were tired or absentminded, I would be prone to make a dangerous mistake. Since expectations play such an important role in supporting conventions – as well as descriptive norms – a change in expectations (of others' conformity) may be another reason why we stop following a convention we observed until now. We may subsequently realize it was a mistake, because the convention is still followed, and revert to the old behavior. Alternatively, when a new convention is in place we are more likely to fluctuate in compliance. Dress codes are a good example. It is now customary in many American companies to have a day (usually Friday) of "business casual" dressing. Many friends reported embarrassing situations in which they were the only ones in jeans and sneakers, only

[31] Even criminal law recognizes mitigating circumstances such as duress, coercion, insanity, and accident.

to realize that the following Friday, when they reverted to dressier suits, more coworkers had adopted the "dress down" code. It usually takes some time to stabilize on a common dress code, and in the meantime behavior can be quite hectic.

The case of social norms is more complex. Norms are sometimes stated in vague and general terms and operate in the presence of areas of indeterminacy and ambiguity. Several norms may apply to the same situation, or it may not be clear which norms have a bearing in a given case. Whenever it is unclear which norm applies to a given situation, we may of course expect irregular behavior, as the former example of tipping in a foreign country illustrates. Variance is also to be expected (or at least it is explainable) when sanctions have been introduced or removed or, for some reason, there has been a change in expectations. With fairness and reciprocity norms, it is often in one's interest to break the norm, to yield to temptation. Why should I accept a fair division if I have the upper hand and, moreover, I will not interact with my partner in the future? Why should I reciprocate my neighbor's favors if I am moving to a different town soon? My sudden transformation can be altogether explained by self-interest, boosted by a change in sanctions and expectations. Another possible reason for inconsistent behavior is weakness of the will. Whenever the temptation is too great, the bait too alluring, I may break a norm that I otherwise approve of and regularly obey.

Yet, if no such reason is apparent and we know that a person (a) approves of a given norm and (b) has conformed to it on other, similar occasions, we could either conclude that norms' influence on behavior has been overstated or that we need a better understanding of the role of situational cues in inducing conformity. Indeed, factors having nothing to do with the norm in question – including other norms, attitudes, or environmental factors – may attenuate or emphasize its impact on actions. Environmental stimuli in particular have been reported by psychologists to cause major changes in the kinds of behaviors, such as the propensity to help other people, that we usually expect to manifest a certain consistency and that are taken to signal a character disposition. Several studies of helping behavior indicate that people are more likely to help others if they are in a familiar environment, or if the request comes from a female. When facing emergencies, people are much more likely to intervene if they are alone. The presence of other bystanders to an accident seems to consistently dampen altruistic ardor (Latane and Darley 1968). Similarly, we have no indication of a general disposition to take normative considerations as overriding, or of an unfailing

inclination to obey a norm whenever a norm is in place. Quite to the contrary, all the evidence we have points to situational factors as having a significant influence on behavior. However, as much as situational factors may attenuate the impact of norms on behavior, the opposite is also true: Situational factors may increase the effect of norms on behavior by making a norm salient. Unfortunately, there are no experiments tracking personal (as opposed to interpersonal) variations in behavior in similar situations, where the experimenter slightly varies the environment or the description of the situation. In the following chapters, I will present some indirect evidence that supports the hypothesis that situational variables are extremely important in focusing actors on social norms, thus inducing or preventing conformity.

Awareness and Choice

In the next chapters, the idea that norms influence behavior only when they are salient or focal for the individual at the time of behavior will be expanded on and put to the test. If people are not strongly focused on a norm, I shall argue, even strong personal norms are not predictive of relevant behavior. Normative focus, in turn, is enhanced or mitigated by situational cues that draw attention to (or distract attention from) a relevant norm. There is by now a large database of experimental results from Trust, Ultimatum, and Social Dilemma games, in which small alterations in the environment or the way in which the game is presented produce major behavioral changes. Individuals may be cooperative on some occasions and selfish in others, give generously or reciprocate at times and be 'mean' at other times. If a fairness norm is activated in condition *x*, when the game is one-shot and the players anonymous, why is it not activated in the slightly different condition *y*, in a similar one-shot, anonymous encounter? Because the apparently inconsistent behaviors are not correlated with the presence or absence of sanctions, this variability has led several authors to discount the importance of norms as explanatory variables in such experiments (Dawes et al. 1977). The reasoning leading to this conclusion is that – if a person were to uphold a norm – then that person would conform to it in all circumstances to which the norm applies. This belief presupposes that (a) we are always aware of our personal standards and ready to act on them, and (b) situational factors have no influence on our behavioral dispositions. Because I will focus next on situational factors and their influence, I will now restrict my attention to the issue of normativity and choice. For example, when situational factors

are paramount, in the sense that their presence is crucial in priming a norm, does it make sense to say that a person *chooses* to follow a norm? If one is unaware of the stimuli and the cognitive process whose outcome is norm-congruent behavior, can we still claim that it is *rational* to follow that norm?

When mentioning the expectations and preferences that support conformity to a social norm, I referred to *reasons* for following a norm, and having reasons can be interpreted as mentally referring to a norm before acting, having intentions, and making a reasoned (and rational) choice. For example, we may say that the trader who uses the conventional signaling system is making a rational choice, because we assume she wants to communicate and, through communication, reach her goal of buying and selling shares. There is a difference, though, between choosing rationally and choosing a course of action *because* it is the rational thing to do. In light of the coordinating role played by the trading-signaling system, and assuming the trader's goal is to make trades, we judge the trader's choice to be rational, but the trader herself may have been totally unaware of having a choice. In this case, what has been activated is not the deliberational route to behavior but rather the heuristic one. The trader may have never thought about the signaling convention being a coordinating device, nor might she be aware of any goal or plan that following the convention helps her to achieve. This, I must add, is a common experience; frequently we do not think much before acting, in the sense that our behavior does not consciously follow from intentions or plans and is carried out without awareness or attention. To engage in thoughtful processes, we must be sufficiently motivated: The situation must have high personal relevance, our action must have important consequences, we are held responsible for our choice, or there is some challenge present. As opposed to this thoughtful evaluation of pros and cons, we usually engage in a more rapid, heuristic form of processing. The trader uses the signaling convention as a default, without a thought to the benefits her behavior yields.

Even obeying a social norm can be, though by no means has to be, an entirely automatic affair. We are, so to speak, in the grip of the situation that primed the norm and are following it through the heuristic route. Those individuals who cooperate in the initial stages of an experimental, finitely repeated public good game do not seem to have gone through a mental process in which they calculated the costs and benefits of being nice. Indeed, a simple calculation of costs and benefits might have induced them to defect immediately, as game theory predicts they

will do. On the other hand, these people are not dupes: Cooperation precipitously decays whenever people realize they have been cheated by others (Dawes and Thaler 1988; Fehr and Gachter 2000b). My hypothesis is that subjects in experiments act like any of us would in a new situation and use social norms as defaults, at least initially. If not challenged, a cooperative norm is adopted in all those situations in which it is made focal. If, however, the norm is violated often enough to be noticed, people will stop following it, at least in that situation. Recall that my definition of social norms entails that an individual needs to have conditional preferences and the right kind of expectations in order to follow a norm. The potential norm follower was represented as facing a Bayesian game. If he initially assesses a higher probability to being matched with another norm follower, he will behave cooperatively. But he will revert to defecting if he realizes his expectations are not met. I am not claiming here that mine is a realistic model of how we reason, but, as will be made plain in the following chapters, I maintain it is a fairly good explanatory and predictive model, because my definitions are operational and their consequences are testable. Furthermore, the fact that we are not aware of our mental processes does not mean that the beliefs and preferences that underlie the choice to conform have no existence. On many occasions our conscious awareness of a norm, and of the expectations and preferences that trigger conformity, is only brought about by the realization that the norm has been violated.

Suppose you are one of the nice guys who choose to cooperate in a finitely repeated public good game. When asked to explain your behavior, you may offer a rational justification and refer to the choice to obey a norm: You may say that you would really feel guilty not to give it a chance and signal your good intentions. Or you may say that being cooperative is a good rule, and that it is better, in the long run, than being a defector, and therefore you are committed to it even on those occasions in which you may cheat with impunity. Your rational justification is part of a narrative, an acceptable account of why we act as we do. Cognitive psychologists tell us that we often have little direct introspective awareness or access to our higher level cognitive processes (Nisbett and Wilson 1977).[32] We may be unaware that certain stimuli influence our responses, or we may even be unaware of the existence of stimuli that have a causal effect on our responses. Yet when questioned about our choices, judgments, and

[32] A high-order cognitive process mediates the effects of a stimulus on a complex response such as judgment, inference, problem solving, and choice.

evaluations, we are usually quite articulate in offering credible reasons. A plausible explanation is that our reports are based on implicit theories about the causal connection between stimulus and response. The causal theory we put forth may happen to be an accurate account of what stimulus was influential in producing our response, but accuracy, according to Nisbett and Wilson, is not synonymous with awareness. We may accurately report that a particular stimulus was influential in producing a behavioral response because the stimulus is available and salient, and it appears to be a plausible cause, not because we have a privileged access to our higher cognitive processes. If the actual stimulus is not available, salient, or not deemed to be a plausible cause of the response, it will regularly be discounted as uninfluential.[33]

Latane and Darley's (1968) experiments on helping behavior offer a disturbing example of how choices may be influenced by factors that are outside our immediate awareness. Their subjects were progressively more unlikely to help somebody in distress as the number of bystanders increased, but they were entirely unaware of the effect that the presence of other people had on their behavior. Moreover, when the experiments were described in detail to different, nonparticipating subjects, who were then asked to predict how others (and perhaps themselves) would behave in similar circumstances, they concurred that the presence of other people would have no effect on helping behavior. In this as well as other similar experiments, the congruence between the participants' reports and the predictions made by nonparticipants suggests that both are drawn from a similar source. Nisbett and Wilson explain the congruence by referring to common, shared causal theories that make both actor and observer 'perceive' covariations between particular stimuli and responses.[34]

Some of our reports may instead be highly accurate, as when we apply the sequential steps of a decision process we have learned. A business school graduate who is making the decision whether to buy a particular stock, for example, will apply learned rules for evaluating the stock and weighing all the factors that have a bearing on its price. Her report on

[33] We are usually blind to contextual factors, as well as to position, serial order, and anchoring effects. Most people would think it is outrageous that the choices they make might be influenced by such irrelevant factors as the position (say, from left to right) of the object chosen.

[34] The criterion for awareness proposed is a "verbal report which exceeds in accuracy that obtained from observers provided with a general description of the stimulus and response in question" (Nisbett and Wilson 1977, p. 251).

her final choice will accurately list the weighted factors as reasons for her choice. Similarly, we might be fairly accurate about the weights we assign to various factors in deciding what a fair division of a particular good should be. But this may happen because our culture (or subculture) specifies rather clearly which factors should count in such a decision. Still, being able to describe the evaluative criteria one has applied is not evidence of direct access to one's mental evaluation process.

The existence of a norm and of reasons for conformity might thus be correctly reported as an explanation for our behavior, even if we are unaware of the complex mental process that resulted in that behavior. Situational dependency can in turn be understood in two different ways. One is that the environment or situation we are in provides perceptual stimuli to which we respond in an 'automatic,' unreflective way. *Ex post*, we may or may not accurately report on the importance of the stimuli, depending on whether they are available and how plausible they are as causal factors. Alternatively, we may see the situation as influencing the way in which we consciously interpret and understand our surroundings. A norm in this case can be made salient by particular situational cues, but we still *choose* to follow it, that is, consider alternatives and make mental reference to the norm before we act. I believe both accounts of situational dependency to be valid, depending on the level of awareness we experience at any given time. There are occasions in which we are unaware of the reasons why we do what we do, and occasions in which we are consciously thinking of a norm, and the reasons for following it, before acting.[35] Also in this second case, though, we should not confuse access to our private store of knowledge, emotions, or plans with access to our cognitive processes, which are opaque to introspection.[36]

Lack of awareness should not be equated with lack of rationality. It is possible to maintain that it is rational to follow a norm, even if for the most part our subjective experience of conformity to a norm is beyond

[35] A mental state is *conscious* when it is accompanied by a roughly simultaneous, higher order thought about that very mental state. For example, a conscious experience of pain involves more than the simple registering of a painful sensation in the mind. It also includes a realization that one is having this sensation, a thought that "I am feeling pain."

[36] Jones and Nisbett (1972) distinguish between *content* and *process*. Content includes all sorts of private knowledge we possess: We know personal historical facts, our focus of attention at any given time, what we feel and sense, our evaluations, and our plans. They convincingly maintain that we have introspective access to content but not to mental processes.

rational calculation. Compliance may look like a habit, thoughtless and automatic, or it may be guided by feelings of anxiety at the thought of what might happen if one violates the norm. Yet conformity to a norm may be rational, and may be explained by the agents' beliefs and desires, even though one does not conform out of a conscious rational calculation. As David Lewis himself pointed out in his analysis of habits, a habit may be under an agent's rational control in the following sense: If that habit ever ceased to serve the agent's desires according to his beliefs, it would at once be overridden and abandoned.[37] Similarly, an explanation in terms of norms does not compete with one in terms of expectations and preferences, because a norm persists precisely because of certain expectations and preferences: If I ever wanted to be different, or if I expected others to do something different, I would probably overcome the force of the norm.

We may conclude that awareness is not a necessary condition for being rational, in the sense that, even if unaware, we may still act according to our beliefs and desires. To maintain that following a norm can be described, at least in principle, in terms of beliefs and desires and hence as a (practically) rational choice allows us to think of norms as a special kind of unintended collective outcome of individual choices.[38] Such outcomes have desirable properties, for example, they are equilibria of coordination games. Note that being an equilibrium does not make a social norm good or efficient; there are lots of bad equilibria around. It simply means that the expectations and actions of all the parties concerned are consistent, or that their expectations are self-fulfilling. This raises the important question of how such consistency comes about, but we will discuss this later. Another important advantage of defining norms in terms of beliefs and preferences is that we are providing an operational definition of what a norm is. This is important in experimental studies, where we want to assess whether the behavior we observe is due to the presence of norms or to something else. If we know that norms are only followed if certain expectations exist, then it is possible to verify if indeed people have those expectations, or to manipulate them in order to see whether their behavior changes in predictable ways.

[37] D. Lewis, 1975, p. 25.

[38] It is important to distinguish between *practical* and *epistemic* rationality (Bicchieri 1993). Practical rationality is the rationality of an action, given the agent's goal and beliefs. Thus goals may be unrealistic and beliefs false, and an action still may be practically rational. Conversely, epistemic rationality is the rationality of the beliefs we hold.

Appendix to Chapter 1

In this short appendix I introduce a general utility function based on norms. Consider a typical n-person (normal-form) game. For ease of formal treatment, think of a norm as a function that maps one's expectations concerning the behavior of others into what one "ought to do." In other words, a norm regulates behavior conditional on other people's behavior.

Denote the strategy set of player i by S_i, and let $S_{-i} = \Pi_{j \neq i} S_j$ be the set of strategy profiles of players other than i. Then a norm for player i is formally represented by a function N_i: $L_{-i} \rightarrow S_i$, where $L_{-i} \subseteq S_{-i}$.[39] In an n-person Prisoner's Dilemma game, for example, a shared norm may be to cooperate. In that case, L_{-i} includes all the strategies of all players (excluding player i) that prescribe cooperation.

Two features of this definition are worth noting. First, given the other players' strategies, there may or may not be a norm that prescribes how player i ought to behave. So L_{-i} need not be, and usually is not, equal to S_{-i}. In particular, L_{-i} could be empty in the situation where there is no norm whatsoever to regulate player i's behavior. Second, there could be norms that regulate joint behaviors. A norm, for example, that regulates the joint behaviors of players i and j may be represented by $N_{i,j}$: $L_{-i,-j} \rightarrow S_i \times S_j$, where $L_{-i,-j}$ is the set of strategies adopted by all players other than i and j. Because I am primarily concerned with two-person games, I will not further complicate the model in that direction.

A strategy profile $s = (s_1, \ldots, s_n)$ *instantiates* a norm for j if $s_{-j} \in L_{-j}$, that is, if N_j is defined at s_{-j}. It *violates* a norm if, for some j, it instantiates a norm for j but $s_j \neq N_j(s_{-j})$. Let π_i be the payoff function of player i. The norm-based utility function of player i depends on the strategy profile s and is given by

$$U_i(s) = \pi_i(s) - k_i \max_{s_{-j} \in L_{-j}} \max_{m \neq j} \{\pi_m(s_{-j}, N_j(s_{-j})) - \pi_m(s), 0\},$$

where $k_i \geq 0$ is a constant representing a player's sensitivity to the relevant norm.[40] The first maximum operator takes care of the possibility that the norm instantiation (and violation) might be ambiguous, in the sense that a strategy profile instantiates a norm for several players simultaneously. However, this situation never occurs in my examples, so the first maximum

[39] Note that N need not be deterministic. As we shall see in Chapter 3, when we look at Ultimatum games, N can also be a random variable.

[40] k_i is only unique up to some positive factor that varies according to the players' payoff functions.

operator degenerates. The second maximum operator ranges over all the players other than the norm violator. In plain words, the discounting term (multiplied by k_i) is the maximum payoff deduction resulting from all norm violations.

As an example to illustrate the above norm-based utility function, consider the Prisoner's Dilemma, where each player has two possible strategies: C (Cooperate) and D (Defect). The norm-based function for either player is defined at C and undefined at D. The utility function for player 1 is then the following:

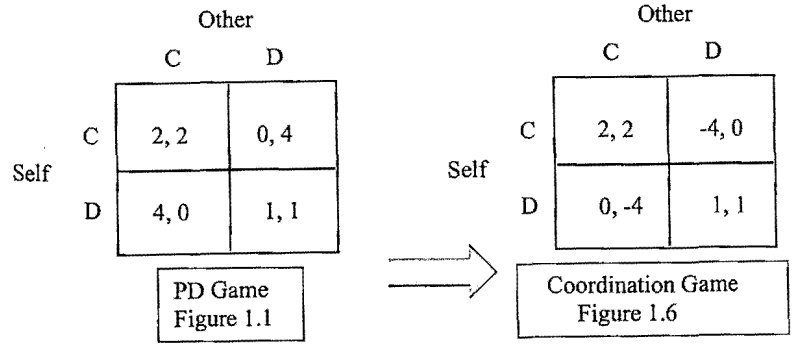
$$\begin{aligned}
U_1(C, C) &= \pi_1(C, C) - k_1(\pi_1(C, C) - \pi_1(C, C)) = \pi_1(C, C) \\
U_1(D, D) &= \pi_1(D, D) - k_1(\pi_1(D, D) - \pi_1(D, D)) = \pi_1(D, D) \\
U_1(C, D) &= \pi_1(C, D) - k_1(\pi_1(C, C) - \pi_1(C, D)) \\
U_1(D, C) &= \pi_1(D, C) - k_1(\pi_2(C, C) - \pi_2(D, C)).
\end{aligned}$$

Player 2's utility function is similar. The game turns out to be a coordination game with two equilibria when $U_1(D, C) < U_1(C, C)$ and $U_2(C, D) < U_2(C, C)$, that is, when[41]

$$k_1 > \frac{\pi_1(D, C) - \pi_1(C, C)}{\pi_2(C, C) - \pi_2(D, C)}$$
$$k_2 > \frac{\pi_2(C, D) - \pi_2(C, C)}{\pi_1(C, C) - \pi_1(C, D)}.$$

Otherwise it remains a PD game.

As an example, take the PD game in Figure 1.1 and assume the players' payoffs are as follows:



[41] Note that $U_1(D,C)$ stands for the utility of player 1 when 1 plays D and 2 plays C. Analogously, $U_2(D,C)$ stands for the utility of player 2 when 1 plays D and 2 plays C.

In this case, $\pi(C, C) = 2$ and $\pi(D, D) = 1$.
However,

$$U_1(C, D) = 0 - k_1 \max \begin{Bmatrix} \pi_1(C, C) - \pi_1(C, D) \\ \pi_2(C, C) - \pi_2(C, D) \\ 0 \end{Bmatrix} = 0 - k_1(2)$$

and

$$U_1(D, C) = 4 - k_1 \max \begin{Bmatrix} \pi_1(C, C) - \pi_1(D, C) \\ \pi_2(C, C) - \pi_2(D, C) \\ 0 \end{Bmatrix} = 4 - k_1(2);$$

similar calculations hold for player 2.

For both players to prefer to cooperate with each other, it must be that both k_1 and k_2 are greater than 1. For example, if we assume that, say, both k_1 and k_2 are equal to 2, we obtain the above coordination game (Figure 1.6). Note that it is not necessary to assume that k_1 and k_2 are the same. In fact, players may have different degrees of 'sensitivity' to a norm. Being 'sensitive' to a norm simply means that one dislikes being the victim of a norm violation as well as being the transgressor. We may thus say that k defines different types of players. In our simple example, there can be only two types of players: Either a player's k is greater than 1, or it is equal to or less than 1.

In this case, player i (with $k_i > 1$) is rational iff she chooses a strategy s_i such that the expected utility $\mathrm{EU}(s_i) \geq \mathrm{EU}(s_i')$ for all $s_i' \neq s_i \in S_i$, calculated with respect to the probability that $(k_j > 1)$. It is important to remember that when a player is faced with a PD game and has no information about the identity or past actions of the other player, she will rationally choose to 'follow the cooperative norm' if *two* conditions are satisfied. She must be a potential norm-follower (i.e., her k must be greater than 1) and she must believe that the other player's k-value is such that it makes him sensitive to the norm (in our example, it must also be greater than 1). In other words, a norm-follower faced with a PD game will have to assess the probability that the other player is the norm-following type. In our case, if $p(k_2 > 1) > 1/2$, player 1 will choose to cooperate.