To appear in Philosophy of Social Science, N. Cartwright and E. Montuschi, eds., Oxford University Press

Norms, conventions and the power of expectations

Cristina Bicchieri

What is the difference between a chair and a social norm? Both are human artifacts, existing for human use. Yet think for a moment what would happen if, like in an old episode of Twilight Zone, all life on earth was wiped out. All life but one: you alone remain, wandering around in a world now horribly silent. You stumble on a broken chair, unusable. You may not look at it ever again, you will never sit on it, and will soon forget about that broken chair, but until time wipes it out, that chair will exist in its corner of the world. In your previous life, you were a student, a family member, a friend. Each group had its own norms, and some, like reciprocity or truth telling, were very general, spanning across all groups. You were a norm follower; not always, but most of the time. Now there are no norms to speak of. Truth telling makes no sense if there is nobody to talk to, and so it goes for every other norm you can think of. To exist at all, norms need people who collectively believe they exist. You suddenly realize you have followed norms because you thought other people were following them, and also trusted that all those people believed that everyone should obey those norms. You thought of norms as having an independent existence not unlike that broken chair, an existence in the world beyond

what people thought and believed about them. You were wrong: these beliefs are no more, and all norms have therefore ceased to exist.

What is the point of being honest and trustworthy, cooperative and fair, if there is nobody to appreciate and reciprocate it, not to mention to interact with? You used to be the proud member of the "Red bandanna group", a group of motorcycle enthusiasts who met every Sunday to compete with other groups in reckless speed races. That red bandanna was a sign of distinction and belonging, and your pals would have taken offence seeing you without it. They all wore it, and thought it should be very visible on the head of every group member, to signal the privilege of being part of a gang that had won innumerable races. That group is no more, gone are the motorcycles and the red bandannas. You still have one in your pocket, though, should you wear it? You might, if you get nostalgic, but it signaling power is gone. There is nobody to signal anything to, and nobody will get offended and reproach you if you stop wearing it.

Norms are social constructs, like tables and chairs, but much less permanent and independent of our thinking about them. Though norms are expressed in prescribed or proscribed behaviors, actions we can observe or at least describe, these actions would be senseless, or at best lose their original meaning, without the collective beliefs that support them. So the most important question to ask about norms is what system of beliefs supports and defines norms. Once we understand these beliefs, we can tell whether the behaviors that we observe are norm-driven or not, measure the consistency between beliefs and behavior under different conditions, and make predictions about future behaviors. Before we come to define the kind of beliefs that support and define

social norms, however, we shall briefly look at the most common, ubiquitous definitions of social norms, their advantages and shortcomings, and what can we learn from them.

Social norms have been extensively studied in the social sciences, though different disciplines have stressed different features of norms. Yet a shared, common understanding of what a social norm is can be traced across all fields: *a norm refers to a behavior that is collectively approved or disapproved in a group or population, and is enforced by sanctions*. It is also tacitly implied that social norms are very different from moral or legal norms, though this difference is not often articulated in detail. Although legal norms in particular seem to fit the collective approval/disapproval/sanctioning description, there are important differences between legal and social norms. First and foremost, a legal norm is an explicit, mandatory rule of behavior formally established by the state. It usually proscribes behavior, whereas social norms often also prescribe. Social norms are often unspoken and informal, and their origin is not clearly identifiable with a particular moment in time. They have evolved out of protracted social interactions, but we cannot usually tell exactly how they have evolved and in which circumstances. Whereas a legal norm explicitly indicates the conditions of its implementation, the subjects whose actions it regulates, their mutual rights and duties and the sanctions for failure to obey one's duty, a social norm is much less specific. This difference is particularly evident in employer-employee relations. Such relations are formally regulated by contracts, but the greatest influence is usually exerted by non-legal incentives and sanctions, such as reputational concerns and relationship-specific advantages. The social norms that regulate work relations are much less specific than

legal contracts, and the sanctions for non-compliance, such as blacklisting or negative gossip, are entirely informal.

Social norms often engender expectations of compliance that are felt to be legitimate, and close in a sense to 'having a right' to expect certain behaviors on the part of others, who therefore are perceived as 'having an obligation' to act in specific ways. This is because we have an ingrained tendency to move from *what is* to *what ought to be*, and conclude that 'what is' must be right or good. Yet, apart from our longstanding habits of performing and expecting others to perform certain actions, there is no deeper foundation to these presumed 'rights and obligations', however intensely felt they might be. Whereas violation of legal norms elicits formal negative sanctions, and there is no formal reward for complying with the law, the positive and negative sanctions that attach to social norms are quite different. With a social norm, the approved behavior is buoyed up by informal positive sanctions (tangible rewards, praise, status, reputation, etc.) and the censured behavior is discouraged by means of negative sanctions (punishments, shaming, ostracism, ridicule, etc.).

Note that informal sanctions are also used to discourage or support moral norms. It is debatable whether we can draw a sharp distinction between social and moral norms, but we usually refer to norms as 'moral' when they have a universal content -- such as norms against harming others without reason -- and when our allegiance to them tends to be independent of what others do. As we shall see later, a distinguishing feature of social norms is that they are conditionally followed, whereas a moral norm is unconditional. This does not mean that we always obey moral norms. We may be tempted to perform 'immoral' acts, but an excuse such as "I did it because others did it" is

not deemed to be acceptable. In the case of a social norm instead, it is a perfectly reasonable justification. So it is reasonable to state "I did not pay my dues because I know nobody pays", but few would accept "I raped and killed the prisoners because all my fellow soldiers did it" as a good reason to perform such a horrific act.

Sometime social norms can get internalized to the extent that they do not need social enforcement and are spontaneously adhered to by individuals. In this case we say that individuals are directly motivated to comply with the norm, whereas the use of sanctions is a form of indirect motivation. Sanctions, however, keep playing a role, since "direct motivation" may be tied to feelings of guilt and shame that the mere thought of transgression evokes. In this case an internal monitoring mechanism has taken the place of social monitoring and sanctioning. The sociologist Talcott Parsons (1951) was one of the main proponents of norm internalization. In his view, we form lasting dispositions to conform to our society's norms through a process of socialization that starts within the family. This view has been criticized on two counts. The first refers to the long-standing debate between methodological individualism and holism, which is also discussed in chapters X (on institutions) and Y (on game theory) of this volume. Since the Parsonian view accords priority to social value systems and views individuals as bearers of social values, individual actors can no longer be the basic units of analysis. This has major consequences for the study of social institutions. In many a sociologist's view, social institutions cannot be explained as resulting from individual actions and interactions, since actions are not a primitive unit, independent of those very institutions they are supposed to explain. Institutions and their relations are the primitive units of analysis, irreducible to the microsphere of the individuals that act within their scope. This stands

in sharp contrast to the individualist's view, according to which all social institutions can be explained as resulting from, and being reducible to, individual agency.

The second and more damaging criticism of Parson's view of internalization has to do with the empirical adequacy of what can be inferred from it. Norms that are internalized will be very resistant to change, and we should observe a positive correlation between *personal* normative beliefs and action. Yet history is full of examples of rapid norm change (think of smoking and sexual mores), as well as swift norm emergence. As to the positive expected correlation between personal normative beliefs and behavior, there are many studies in social psychology that fail to find it. Even personal normative beliefs that are typically acquired during childhood, such as honesty, tend to be uncorrelated with behavior. The more secure positive correlation that is regularly observed is one between *social* normative beliefs and behavior. Whenever individuals think that the relevant group holds certain normative beliefs, they will be inclined to act according to them even if, personally, they may be indifferent or even opposed to those actions . Indeed, the existence of a social norm usually leads to actions consistent with it, provided it demands behavior that a reference group defines as appropriate and desirable; this fact, however, militates against the internalization view of norms as a *general* account of conformity.

There are, however, two alternative interpretations of internalization that do not stand in sharp contrast with these empirical observations about usual norm compliance. One interpretation is moral, the other cognitive. We cannot deny that there exist norms that we have internalized to the point that almost no variance exists in norm-induced behaviors. Such norms are typically proscriptive, and as such not likely to be correlated

with observable behavior. Take, for example, norms that proscribe inflicting unwarranted, gratuitous harm. By and large, harming someone for no reason is not even conceived as an option, as the mere thought of performing a destructive, unjustified act spawns revulsion and guilt in most of us. Such norms are internalized to the point that we become aware of their force only when faced with a violation, and our allegiance is perceived as unconditional. What we usually call moral norms, insofar as we understand them to be internalized, unconditional imperatives, fit the above description. It is an open question whether we are born with a 'moral organ' shaped by evolution or it is society that shapes our moral sensibilities. In any case, social norms are not unconditionally followed and are not internalized in the above sense, unless we want to blunt the boundaries between the moral and the social.

The second interpretation of internalization is my own cognitive one. Many of the norms we follow are *learned*, often through repeated interactions in a variety of situations that we come to categorize as typical cases to which the norm applies. So we learn to accept and return favors and gifts, but not bribes; to share equally, but also reward merit and make allowance for need. We learn when and how to greet, how to behave at a party and what to say at a funeral. When we find ourselves in one 'typical' situation where a learned norm applies, we tend to conform in automatic ways. The norms we learn we uphold as "default rules", ready to apply them to similar cases until it becomes evident that conformity has become too costly. It is well known that, in repeated social dilemmas of the kind discussed in Chapter X [KSs] , players start by cooperating but cooperation precipitously declines as soon as someone; the players adopt cooperation as a default rule, but are ready to abandon it when they realize it has

significant costs. Internalization in this sense means we economize on thought, not that the norms that society has imposed on us are so deeply entrenched as to be inflexible and unchangeable.

The very generic definition of norms as socially approved behavior supported by sanctions tells us two important things: first, social norms are closely tied to *sanctions*; without sanctions (internal or external), they may not exist. Second, part of the very definition of norm is that it refers to behaviors and patterns of behavior that are collectively approved or disapproved, hence such behaviors *matter* to people. A lot of attention has been paid to the question why certain behaviors matter so much that people will go to great lengths to make sure they are adhered to, and engage in costly sanctioning to support them. An obvious answer is that norms perform critical social *functions*, such as attaining social order and useful collective action, or even help one group to exclude or discriminate against another group, thus keeping vital resources within the group . Yet saying that a norm performs an important social function does not explain how it originated, or why we keep obeying it. A norm may have evolved to smooth social interactions, and it may keep doing so quite efficiently, but we would be hard pressed to say that this is the reason why it came about.

Take reciprocity: it is certainly important to live in a social environment where people reciprocate valuable, beneficial actions. Without reciprocity there would be no trust, and without trust we would have no markets or modern political systems: markets and democracy rely on people trusting their business partners, as well as their elected representatives. Recognizing the social importance of trust, we care for and support norms of reciprocity, and in this sense we may say that their *stability* is linked to the

social functions they fulfill. Their *origin*, however, cannot be explained in this way. A social function a norm comes to play is not the cause of its spontaneous emergence. Take, again, a norm of reciprocity. A society may have evolved *several* strategies that promote reciprocation of trust; all these strategies involve some punishment for non-reciprocators, from mild to harsh, and they can exist along each other. Taken together, all these strategies result in observationally equivalent behaviors: almost all individuals trust and reciprocate, but the norm itself results from many different strategies. The individuals who adopt one of these strategies do so because it is in their long-run interest to reciprocate trust, not because society at large benefits from it.

Yet the social function a norm plays may, and often does, explain its stability within a population. This statement needs qualification. A norm is stable if it is durably obeyed by great part of the population (or group) in which the norm exists. Transgressions occur, but they do not challenge the norm's permanence. This does not mean that compliance with a norm is mainly due to our being conscious of the social functions it performs, so that knowing its beneficial function gives us an overriding *reason* to obey it. Often we are not fully aware of the social benefits of a particular norm, but even if we are, our reasons to obey it are often much less worthy. If knowing the benefits of cooperation were a sufficient reason to cooperate, there would be no free riders, i.e., people who do not cooperate with others but benefit from the fact that others cooperate. Think of taxpayers; we have public services that anyone can use because people support them by paying taxes. Someone who does not pay, however, can still use and benefit from these services. We have norms precisely because there is often a tension

between what we would like to do (skirt a common task, avoid paying our dues, being less than fair) and what is socially beneficial.

To explain a norm's stability, as well as its existence, we have to look further into the reasons why individuals conform. If we are less than fully aware of the benefits a norm may bestow upon society or, even if aware, we are not fully motivated by them, why do we conform? Many believe the answer lies in the existence of sanctions. They may be internal, as when we say that a norm has been *internalized*, by which we mean it has become part of our value system. Or, more often than not, external sanctions are at work to keep people in line. In this case, we say that there is a *rational* motivation to obey a norm: we want to reap the benefits of conformity or avoid the costs of transgression. If we think of individuals as rational decision-makers, we can see the appeal of this view of norm compliance. Norms are exogenous, external constraints we have to take into account when we make a choice. Economists are particularly fond of this view; a norm is, not unlike a budget constraint, a constraint on the set of possible actions one may take. The constraint, in fact, is not the norm per se, but the expected consequences of disobeying it. In this view, if the expected benefits of transgressing a norm are greater than the expected punishment, people will not conform. A main point to notice about the cost/benefit view of conformity is that it disentangles conformity from attribution of value, i.e., one may conform to a norm even if, for that individual, the norm has no value. We shall come back to this important point later.

Yet most of our actions do not stem from a cost/benefit calculation, at least not a conscious one. As I mentioned before, norms are more often than not like default rules that we mindlessly follow in the appropriate circumstances. We are not *aware* of the

possible sanctions, even if it is hard to deny that sanctions often do play a role in driving conscious compliance, especially in those cases in which people do not care much about what the norm stands for (think of foreign women having to cover their heads in a Muslim country). Since the cost/benefit model can be thought of as a *rational reconstruction* of norm conformity, we can disentangle awareness from rational choice. The cost/benefit model specifies when and why norm compliance is rational, but it does not profess to be a realistic, precise description of the way we in fact deliberate. It just says that, were we aware of the presence of sanctions, we would choose the course of action that minimizes costs. Even a rational reconstruction, however, has its constraints. A cost/benefit model must require that, were sanctions clearly absent, behavior would change in predictable ways. A good model, however abstract, must make testable predictions.

Another, more important objection to the cost/benefit view of motivation has to do with interdependent expectations. In that view, avoidance of negative sanctions constitutes a decisive reason to conform, *irrespective of what others do*. The only expectations that matter are those about the sanctions that will ensue. What others do or do not do is irrelevant to motivation. The traditional rational choice/cost-benefit model depicts a decision-maker that stands alone in the face of uncertainty and a measurement problem. All that matters is how one should assess the present and future costs and benefits of incurring or avoiding sanctions, the severity of these sanctions, and the probability of being monitored and caught.

In reality we are embedded in a thick network of relations, we constantly interact with others, and what they do and think matter a great deal to us. Game theory that, as

we see in Chapter X (decision and game theory), studies interactive decision-making, provides a framework for understanding social interactions and the mutual expectations that accompany them and thus provides good, if incomplete, models for the kind of decisions involved in following a norm. It allows a micro-level analysis of how the incentives to behave in specific ways are influenced by others' behavior, and in what way behaviors are interdependent.

From social psychology we know that only *some* expectations are positively correlated with norm-abiding behavior: only those normative beliefs that people perceive to be collectively shared and put into practice matter to action. Casting aside for the moment the issue of how to differentiate among types of expectations, we want to know how a game-theoretic model can broadly represent how individuals' expectations converge and prompt individuals to behave according to them. To see what I mean by convergence of expectations, let's suppose we all believe, for whatever reason, that within a month there will be a market crash. We act on those beliefs and immediately sell our stock positions. This sudden sale depresses stock prices and, indeed, the market tanks. The market expectations were by no means normative, but they give us a vivid example of how collective expectations can bring on actions that make those expectations true. These are what we call *self-fulfilling expectations* and I want to argue that they have a lot to do with how norms persist. As you learn from Ch X [KSs], game theory is a good tool if we want to model the interaction of beliefs (expectations) and behavior, and several authors have used it to give an account of norms and conventions. However, as we shall see, game theory falls short of producing a satisfactory account of the *role* different kinds of expectation play in conformity.

Norms as equilibria: the game theory connection

Various authors, including myself, have proposed a game-theoretic account according to which a convention or a norm are broadly defined as *Nash equilibria*. As Chapter X (decision and game theory) explains, a Nash equilibrium is a combination of strategies, one for each player, such that each individual's strategy is a best reply to the others' strategies, were one to take them as given. This means that a Nash equilibrium is an outcome of the game from which no individual player has any incentive to diverge. This outcome, however, is not necessarily the most efficient (think of the mutual defection equilibrium in the prisoner's dilemma described in Ch X [KSs]); it is just one to which the players will converge, if they are acting rationally on the basis of their beliefs.

Most of these authors were interested in conventions, but social norms, too, can be thought of as equilibria. Since it is an equilibrium, a norm is supported by self-fulfilling expectations, in the sense that in equilibrium players' beliefs are mutually consistent, and thus the actions that follow from those beliefs will validate them. Characterizing social norms as equilibria has the advantage of emphasizing the role that expectations play in upholding norms. On the other hand, this interpretation of social norms does not *prima facie* explain *why* people prefer to conform if they expect others to conform.[1] After all, if everyone cooperates the defector will reap great benefits, so the mere expectation of universal cooperation may not be enough to induce good behavior.

[1] The numbers (or letters) we put in the cells of a game matrix represent 'utiles' that illustrate how much a player likes a particular outcome of the game. *Why* a player has those 'utiles' is not a question game theory can answer.

When I mentioned the stability of norms I said that, in order to explain stability, we have to understand why people conform. The game-theoretic account gives a clear answer in the case of conventions. Take for example conventions such as putting the fork to the left of the plate, adopting a dress code, using a particular sign language, or blowing one's nose with a handkerchief. In all these cases, my choice to follow a certain behavioral rule is usually conditional upon expecting most other people to follow it. I say 'usually' because one may have other, overriding reasons to follow a rule. Take dress codes: For a very religious person, wearing a yarmulke/skullcap may represent one's compact with God, and be completely independent of expecting others to do the same. In this case we cannot say that that person is following a convention. In a convention, on the contrary, mutual expectations are everything. Once my expectation about others' behavior is met, I have every reason to adopt the behavior in question. If I do not use the sign language everybody else uses, I will not be able to communicate, and if I blow my nose in my hands when everybody else uses handkerchiefs, I will send out the wrong signal about who I am. It is in my immediate interest to follow a convention, if my main goal is to *coordinate* with other people. So if I expect others to act in a certain way, and I want to coordinate with them, I will adopt that behavior. *Why* one may want to coordinate with others is another issue. All the examples that are typically employed to illustrate conventions, like driving or language, rely on the idea that one's goal can only be achieved by doing what others do. So if I want to be safe on the road or just communicate, I will have to coordinate my actions with those of other people.

In the case of conventions, there is continuity between individual's self interest and the interests of the community that support the convention. It is an example of

harmonious interests: We all want to drive safely and speak the same language. This is the reason why David Lewis (1969) represented conventions as equilibria of *coordination games*. Such games have multiple equilibria, in the sense that – to communicate – we need a language, but which one among many is irrelevant. (See Ch X [KSs] for more discussion of multiple equilibria.) The same goes for driving: to be safe, we need to all drive to the right or to the left, but which way we coordinate upon is irrelevant. Any of these coordination points is, from the viewpoint of achieving our common goal, an equally plausible outcome. Once one of the possible equilibria has been established, players will have every incentive to keep playing it, as any deviation will be costly for the deviant.

Social norms are a different story. For one, the fact that sanctions do play a role in compliance suggests that following a norm may not be in the individual's immediate interest. Behaviors that are socially beneficial, when not mandated by law, are normally supported by social norms that involve sanctions, both positive and negative. Social dilemmas, such as overpopulation, pollution, or energy conservation are examples of situations in which each individual profits from free riding, but the group is better off if everyone contributes. Pro-social norms such as norms of cooperation or reciprocity have evolved to solve such dilemmas, and we often refer to them as unambiguous examples of the discontinuity between individual and collective interests. Not all norms have evolved for this reason, however. Norms of honor killing do not seem to be related to the provision of any collective good, even if honor is the highest valued virtue in some cultures.[2] Conforming to an honor code confers or restores status to those who comply

[2] Honor killings are murders of women by family members that are justified as removing some imputed stain on the family's honor. Men are occasionally killed, but the large majority of victims are women. In

with it, and not conforming is severely sanctioned by the community. Strict honor codes are costly to enforce (you may have to kill your own sister or daughter), but the temptation to evade the norm is tempered by the presence of positive (status, honor) and negative (stigma, ridicule, lack of trust) sanctions. In [Yotam Feldners' words, "the honor of the Arab family or tribe, the respect accorded it, can be gravely damaged when one of its women's chastity is violated or when her reputation is tainted. Consequently, a violation of a woman's honor requires severe action, as Tarrad Fayiz, a Jordanian tribal leader, explains: 'A woman is like an olive tree. When its branch catches woodworm, it has to be chopped off so that society stays clean and pure.' The murder of women to salvage their family's honor results in good part from the social and psychological pressure felt by the killers, as they explain in their confessions. Murderers repeatedly testify that their immediate social circle, family, clan, village, or others expected them and encouraged them to commit the murder. From society's perspective, refraining from killing the woman debases her relatives."[3]

We still have a tension between the individual and the collective, albeit one that is less transparent than what we see in a social dilemma. The point to be made is that social norms, as opposed to conventions, do not arise out of situations of harmonious interests, but out of situations in which a potential or open conflict exists between individual and group interests.

You will learn about game theory in Chapter X [KSs] and I will presume here that you are comfortable with some of the ideas and language from there. The typical game

---

the honor killing culture, a man who refrains from "washing shame with blood" is a "coward who is not worthy of living ... as less than a man" (Feldner 2000).

[3] Y. Feldner, 2000, p. 42

that represents a state of affairs in which following a pro-social norm would provide a better collective solution than the one attained by a rational, selfish choice, is a *mixed-motive game*. In such games the unique Nash equilibrium represents a suboptimal outcome, but there is no way to do better within the confines of the game. I have argued that pro-social norms, as opposed to conventions, are never born as equilibria of the mixed-motive games they ultimately transform. Whereas a convention is one among several equilibria of a coordination game, a norm can never be an equilibrium of a mixed-motive game (such as, for example, a prisoner's dilemma or a trust game). When a norm exists, however, it *transforms* the original mixed-motive game into a coordination one. As an example, consider the following prisoner's dilemma game (Figure 1), where the payoffs are B=Best, S=Second, T= Third, and W= Worst. Clearly the only Nash equilibrium is for both players to defect (D), in which case both get (T,T), a suboptimal outcome.

Other / Self	Cooperate	Defect
Cooperate	**S, S**	**W, B**
Defect	**B, W**	**T. T**

Prisoner's dilemma

Figure 1

Suppose, however, that society has developed a norm of cooperation: Whenever a social dilemma occurs, it is commonly agreed that the parties should privilege a cooperative attitude. Should, however, does not imply "will", therefore the new game generated by the existence of the cooperative norm has two equilibria: either both players defect or both cooperate (Figure 2).

Other / Self	Cooperate	Defect
Cooperate	**B, B**	**W, T**
Defect	**T, W**	**S, S**

A coordination game

Figure 2

Note that in the new coordination game created by the existence of a cooperative norm, the payoffs are different from those of the original prisoner's dilemma. Now there are two equilibria: If both players follow the cooperative norm they will play the optimal

equilibrium and get (B,B), whereas if they both choose to defect they get (S,S), which is worse than (B,B). Players' payoffs in the new coordination game differ from the original payoffs because their preferences and beliefs will reflect the existence of the norm, which has affected players' incentives. More specifically, if a player knows that a cooperative norm exists and has the right kind of expectations, then she will have a preference to conform to the norm in a situation in which she can choose to cooperate or to defect. In the new game generated by the norm, choosing to defect when others cooperate is not a good choice anymore (T,W). The existence of sanctions for non-compliance explains the lower payoff.

The honor killing norm is quite different. This is not a pro-social norm in the sense cooperation or reciprocity norms are. We do not have a social dilemma to start with, so casting the norm as arising from a mixed-motive game would be a mistake. It does not arise from a coordination problem either, since it is difficult to imagine a situation in which there are many equipossible 'honor codes' people can end up adopting, and what matters to them is just to collectively pick (or stumble upon) one. In cultures in which the reputation and honor of the family are the most important attributes, failure by a member to follow adequate moral conduct weakens the social status of the family. The unwed girls of a shamed family will not find a husband, and men relatives will be scorned and ridiculed. The only way to restore honor and reputation is to 'cut away' what brings shame. Since in these cultures honor is often linked to the 'purity' of women, the duty to restore the lost honor is to 'cut away' the lost woman by killing her.

Yet *when the norm is in place*, we do have a coordination game, since a player can decide to obey/disobey the norm, but in this case, unlike the case of a norm of

cooperation, the honor killing norm may be an *inferior* equilibrium. As I already mentioned, keeping the honor code is costly (you will have to kill your straying daughter/sister or be forever dishonored), and all would benefit from some other arrangement. To see that, let's look at the following matrix, where the payoffs go from B (best) to W (worst) and we assume for simplicity that Others are all choosing the same strategy, i.e., to embrace the honor killing code (H) or to abandon it (Not H):

Others / Self	**All H**	**All Not H**
H	**S, S**	**T, T**
Not H	**W, T**	**B, B**

Figure 3

Observe that for H to be a strict equilibrium, it is necessary that the action of a single player (Self alone chooses to stick to the honor code H) imposes *costs* to all the others who have chosen a different path (Not H). For example, if the whole community has rejected honor killing in favor of a more humane and respectful treatment of women, they will have to punish the deviant as violating human rights, as well as being shaken by the brutality of the act. The worse case scenario for Self is still one in which she flaunts the

honor code when it is still adopted by the group. So Self is punished if he disobeys the norm, as he imposes a cost on the others who instead follow it (W,T) by giving women a bad example of leniency, but Self is also punished if he sticks to the norm when others have changed their ways. These two sanctions will significantly differ. In the disobedience case, the group would ridicule and ostracize Self as well as his family; if instead Self keeps following the (costly) honor code when others have abandoned it, he would bear a significant personal cost without reaping any status benefits.

The two off-diagonal boxes have asymmetric payoffs for Self because I am assuming that the direction of collective change is from honor killing (the original norm) to no honor killing. When a norm is abandoned, the curmudgeon that won't change his ways loses out. The story (and the payoffs) would be very different in case a new norm is built, as people move from no honor killing to honor killing. In this case, the trendsetter who suggests that honor killing is a way of showing family devotion and protecting purity may be offering other people something compelling that induces them to join in. And he would benefit from being in the vanguard and innovative. Unfortunately, the matrix representation is a poor tool to characterize the asymmetry between building and abandoning a norm, because it cannot represent temporal direction.

The simple matrix representation tells us just one thing: any social norm, however generated, *creates* a coordination game of which it is an equilibrium. This simply means that, in the presence of any norm, there are always two possible equilibria: either all follow the norm, or nobody does.[4] Note, however, the profound difference that exists

[4] Note that the simple game theoretic representation given here could also represent situations in which, for a norm to be followed, it is sufficient that a majority follows it. In that case, Self will have to believe that

between norms and conventions in this respect. A convention is characterized as one of many possible equilibria of a coordination game. Once players converge to one of them, deviating has a cost for the deviant, but not for the group. With norms instead, deviations always involve *negative externalities*, which means that the deviation of one individual impacts in a negative way all others. A typical example of multi-person coordination game would be the following:

Others / Self	All play A	All play B	All play C
A	1,1	0, 1-ε	0,1-ε
B	0, 1-ε	1,1	0, 1-ε
C	0, 1-ε	0, 1-ε	1,1

Figure 4

I am assuming here that Others (All) are all playing the same strategy, so that their payoffs are not changed in a perceptible way by Self's deviation.[5] Suppose that A, B and C represent alternative ways of greeting strangers. Strangers being introduced may bow, shake hands, or put the palms together in front of the chest (as in India). Assume all members of a specific population coordinate on one of these greetings. Self is clearly

---

the majority of Others are norm-followers, and that his deviation is not critical, i.e., it will have no effect on the number of followers.

[5] In fact, the change is insignificant and not perceived (it lowers the payoff of everyone else by ε, which is close to zero). However, were many players to deviate, a tipping point may be reached, where the payoff of the convention followers would be diminished in a significant way.

better off by following the existing convention. Doing otherwise may lead to confusion, and being perceived as uncouth and inappropriate. Failing to coordinate with Others is just Self's loss. All the others who follow the convention will not suffer a perceptible loss if Self deviates from it; however, if a greater number of 'deviants' were to be present in the group, the loss may become significant (and perceived as such), as coordination may be lost. This example could represent any convention adopted by a large group. It is in everyone's interest to keep following it, and the lone deviant will pay a price, but she will not be sanctioned by the community, since Others incur no perceptible cost.

A more ambiguous story would be that of two parties who have to coordinate on a particular signaling code. Suppose Self and Other have two possible signaling systems available, Red and Blue, each most preferred by one of them. The following matrix represents this situation, as well as the fact that both players want to communicate and prefer that outcome to following their particular inclinations (i.e., Self will not choose Blue, her most preferred code, unless she is sure Other also chooses it). Clearly *both* players lose by mis-coordinating.

Other / Self	Red	Blue
Red	**1,2**	**0,0**
Blue	**0,0**	**2,1**

Figure 5

Now imagine that the players happen to converge, by trial and error, communication or any other reason, on Red as their common signaling code. A unilateral deviation will still damage *both* players but, depending on the costs imposed on the party that sticks to the convention, the damage may be greater for one of them. For example, if Red is the conventional code, and Other deviates, Self will feel doubly damaged: she is not coordinating with Other *and* she was 'sacrificing' by choosing a code that she did not much like to start with. The matrix would thus look like the following:

Other / Self	Red	Blue
Red	**1, 2**	**-1, 0**
Blue	**0, -1**	**2, 1**

Figure 6

In this case it seems reasonable to assume that some form of sanction might be put in place to discourage even 'innocent mistakes'. For example, the players may decide to impose a monetary penalty on the distracted party.

Other / Self	Red	Blue
Red	**1, 2**	**-1, -2**
Blue	**-2, -1**	**2, 1**

Figure 7

Has the original convention morphed into a norm? Is Other sticking to the Red code because he fears the punishment that will surely follow a deviation, or still chooses Red because, irrespective of the possible punishment, he badly wants to communicate with Self? I think the answer lies in assessing the *reason* a player has for behaving in a particular way. In the case of a norm of cooperation, the shadow of external sanctions may be the main or even the sole reason why one chooses to cooperate, otherwise cooperation would not be in one's best interest. Following an honor killing code, too, may be chiefly motivated by the sanctions that a transgression brings about. Killing one's daughter or sister presumably has a high psychological cost, so one should be 'pushed' to do it.[6] In a convention, which is usually followed by many people (as in Figure 3), the main reason to adhere to it is the desire to coordinate with others in order to fulfill one's goals, which happen to coincide with others' goals. The only sanction that matters is one's failure to coordinate, which presumably has a cost. Even in a two-person convention, when the aim is mutual coordination, there is no tension between what one wants to do and what one is expected to do, regardless of the sanctions that the parties may want to impose on each other to discourage reckless deviations. In a convention,

[6] Typically the killing occurs after a period of warnings and is decided by the whole family.

players may experience a form of what is called *moral hazard*: Since I know that you still want to coordinate with me, and that a small 'distraction' will not harm me too much, I may become cavalier in my behavior. To avoid this form of moral hazard, players may want to impose heavier sanctions.

The power of expectations and epistemic traps

A game theoretic account proves too limited to permit a meaningful discrimination in many ambiguous cases. The numbers (or letters) in the cells of the matrix represent 'utiles' or preferences; we model choices as more or less costly or beneficial and rank choices according to the costs/benefits they confer on the decision-maker, but we have no tools to say, as in the ambiguous case of Figure 7, what costs (lack of coordination or external punishment) matter most to the players. Saying that a norm or a convention is an equilibrium just says that, if a player expects conformity, she will have no reason to deviate and, if *everyone* expects conformity, these expectations will be self-fulfilling. Such equilibria are self-perpetuating: the belief that the norm/convention is (almost) universally endorsed generates widespread conformity, and observation of conformity further confirms expectations of universal endorsement. In other words, game theory, while it stresses the importance of interdependence and mutual expectations, does not convey any information about the *nature* of such expectations. Yet it is precisely the types of expectation that guide us that distinguish a norm from a convention.

Let's go back for a moment to the two games of Figure 1 and 2. For a player who is playing a regular prisoner's dilemma, being informed that her partner is an altruist that always cooperates is not sufficient to induce reciprocal cooperation. It will instead strengthen the temptation to defect. But if a norm of cooperation exits, things have changed. Now if one expects the other to cooperate, reciprocal cooperation has become the best choice. Why? Because defecting triggers negative sanctions. The same happens with a norm of honor killing (Figure 3). If one expects others to follow it, and also expects to be punished for not conforming (and be rewarded for compliance), there is every reason to obey. In both cases there is something else, beyond expecting others to conform, that motivates conformity: a *normative* component that is absent in a convention. To understand this important point, let us define more explicitly the two kinds of expectation that support norm compliance:

(a) Empirical expectations: individuals believe that a sufficiently large part of the relevant group/population conforms to the norm and[7]

(b) Normative expectations: individuals believe that a sufficiently large part of the relevant group/population believes they ought to conform to the norm and may sanction behavior.[8]

[7] Note that the "sufficiently large part" clause tells us that universal compliance is not usually needed for a norm to exist. A few transgressions are a fact of life. However, how much deviance is socially tolerable will depend upon the norm in question; small group norms (think of a youth gang's rules), and well-entrenched social norms (think of reciprocating favors) will typically be followed by almost all members of a group or population whereas with new norms or norms that are not deemed to be socially important greater deviance is usually accepted (think of the bride wearing a white dress). Furthermore, as it is usually unclear how many people follow a norm, different individuals may have different beliefs about the size of the group of followers, and may also have different thresholds for what 'sufficiently large' means. What matters to individual conformity is that an individual believes that her threshold has been reached or surpassed.

If having both types of belief buttresses norm compliance, it follows that one may follow a norm in the presence of the relevant expectations, but disregard it in their absence. To be more specific:

(c) *Conditional preference*: individuals will prefer to conform to a social norm on condition of holding the relevant empirical and normative expectations.

Note that conditional preferences for conforming to a norm are different from a preference for what the norm stands for. For example, my reason to engage in honor killing or cooperate in a given situation does not mean I have a general motive to cooperate or kill to save my honor as such. Having conditional preferences also means that, were my expectations to change, my behavior would change, too.

The triad of empirical and normative expectations and conditional preferences is what, in my view, *defines* social norms. It is a richer definition than the game-theoretic one, since it allows for a clearer distinction between norms and conventions based upon which expectations matter to choice. We can now say that a convention is defined by a simpler dyad: empirical expectations and conditional preferences. In order to adopt a signaling code convention, I only need to believe that almost everyone has adopted it. My preference for using that specific code is conditional upon having certain empirical expectations of group compliance and nothing else. On the contrary, my preference for carrying out an act of honor killing depends on believing that this is the customary norm

---

[8] It is important to emphasize that sometimes we obey norms just because we recognize the legitimacy of other's expectations (Sugden 2001). In this case, it is not so much the external sanction that matters, but an internal one. Transgressions are avoided precisely because one feels others have a 'right' to expect a certain kind of behavior, and one has an 'obligation' to fulfill others' expectations.

in my community, that I am expected to perform such an act and that my whole family will be dishonored and ostracized if I do not perform as I should.

It is worthwhile to point out that whereas empirical expectations are first-order beliefs (I believe others will do so and so), normative expectations are *second-order beliefs*: they are beliefs about the beliefs of other members of the collective (I believe others believe I should do so and so). One's personal inclination to support, like or dislike a particular social norm is not the most relevant variable in determining one's allegiance to it. Normative expectations do matter, and they may significantly differ from personal normative beliefs. A personal normative belief that, say, a family should ensure their daughters are married as soon as they reach puberty may agree with the normative expectation about what one's community believes is appropriate behavior, but it also happens that individuals dislike behaviors mandated by a shared norm. When personal beliefs and normative expectations disagree, I predict that normative expectations, not personal normative beliefs, will guide behavior. This is in line with what social psychologists have observed: beliefs that are perceived to be shared by a relevant group will affect action, whereas personal normative beliefs often fail to do so, especially when they deviate from socially held beliefs.

If we come back to the issue of norm stability, we can now see that a social norm is stable insofar as a majority of followers are motivated to conform. Since conforming to a social norm is *conditionally* preferred (otherwise we are dealing with moral norms or values), a norm's stability will be a function of the stability of the expectations that support it. Let us look at a simple case of two different social norms, N_1 and N_2 that are present in a group G. In both cases, we have that:

1. All members of G believe that all other members of G follow N_1 and N_2.
2. All members of G believe that all other members of G believe one ought to follow N_1 and N_2.

However, in the case of N_2, it is not true that "all members of G believe one ought to follow N_2." In fact, a majority of individuals dislike N_2, and do not think for a moment one ought to follow it. Yet they observe compliance, or what they think are the consequences of compliance, and have no reason to believe that those who conform to the norm dislike it as much as they do. So they do not dare speak out or openly transgress, and a norm nobody likes keeps being followed or, if transgressions occur, they will be kept secret. This is a case of what is known as *pluralistic ignorance*, a cognitive state in which each believes her attitudes and preferences are different from those of similarly situated others, even if public behavior is identical (Miller and McFarland 1987). I maintain that the ensuing set of conditions is a fertile ground for pluralistic ignorance:

a) Individuals engage in social comparison with their reference group. We constantly observe what others do and get clues as to appropriate behavior, others' preferences, etc. In the case of norms, we are influenced by the preferences of other group members, but we do not know the true distribution of preferences, which we try to infer from observing their behavior.

b) Others' behavior is observable. If not, the consequences of such behavior are observable. For example, compliance with norms that regulate sexual behavior or other unobservable behaviors can be assessed by observing the presence or absence of the consequences of such behaviors. In the case of norms that prohibit pre-marital sex, teen pregnancies would be a sign that the norm has been flouted.

c) No transparent communication is possible. Because of shared values, religious reasons, or simply the fear of being shunned or ridiculed as a deviant or just different, we do not express views that we think will put us at a disadvantage.

d) It is assumed that, unlike us, others' behavior is consistent with their attitudes and preferences. There are several possible reasons why this might occur. Fear of embarrassment or the desire to fit in are not easy to observe, so we may come to believe that we experience these emotions more strongly than others do. Another possible cause of the self/other discrepancy is what is called the *attribution error*: We tend to overestimate the extent to which others act on private motives (beliefs, preferences) and instead attribute our own behavior to external factors (social pressure in this case).

e) It is inferred that all but us endorse the observed norm. We discount our personal evidence in favor of what we observe and take it at face value.

f) All end up conforming to the public norm, oblivious to the possibility that they are participants in a group dynamic in which all pretend to support the norm, but in fact all dislike it.

In a state of pluralistic ignorance, individuals are caught into an epistemic trap and will keep following a norm they deeply dislike. How long can this last? One may suspect that a norm that is so much disliked would not be stable, since even small shocks to the system of beliefs that support it would lead to its demise. Once the frequency of true beliefs is conveyed to the relevant population, a change would occur. This is only partially true. When actions are strongly interdependent, it is not sufficient to know, and

possibly reach common knowledge, that most group members dislike N_2. Since a norm is supported by normative expectations, the participants must also be sure that its abandonment will not be followed by negative sanctions. People face a double credibility problem: they must believe that the information they receive about the group members' true beliefs is accurate, and they must also believe that everyone else is committed to change their ways. There are many ways to achieve these goals, and there are several examples in the literature of successful change of negative norms by means of information campaigns, public declarations and common pledges (Bicchieri and Mercier 2011).

I have stated that a norm that is beneficial to society is *in principle* more stable than one that is not, or that is even secretly disliked by its followers. Stability, however, is not a *direct* function of the social benefits a norm confers upon its followers. It must be the case that this beneficial function is recognized and expressed in the beliefs of those who conform to the norm, i.e., there must be a shared belief that the norm is valuable for the group that embraces it. Since mutual beliefs support social norms, a norm's stability is a direct function of the stability of those beliefs.

Conclusion

I have presented a view of social norms that, though it encompasses the traditional understanding of what a social norm is (e.g., behavior that is collectively approved or disapproved and is enforced by sanctions) goes well beyond it. Norms exist because of the expectations of those who follow them. These expectations are not just empirical, as

in the case of conventions; they are normative, too, and may include the belief that transgressions will be punished and compliance rewarded. The cost/benefit model of conformity is right in pointing to the importance of sanctions, but its limit is that it does not grasp the importance of mutual expectations. Game theory is a good modeling tool if we want to highlight the interdependence of actions and mutual expectations; yet it does not offer a language specific enough to discriminate between descriptive (empirical) and normative expectations. That distinction is crucial to understand the difference between social norms and other concepts, as normative expectations are part of what motivates compliance with norms. Finally, a definition of norms in terms of conditional preferences and expectations is *operational*, in that it allows us to make predictions about how changes in expectations will trigger behavioral changes, as well as measure norms and our greater or lesser allegiance to them.

Suggestions for further readings

For a general survey of the literature on social norms, you may read Bicchieri and Muldoon (2011). The game-theoretic view of conventions is to be found in Lewis (1969). Parsons (1951) and Coleman (1990) offer the traditional and the modern sociological views of norms, respectively. Bicchieri (2006) presents a theory of social norms that combines game theory and psychology.

References

Bicchieri, C. (2006). *The Grammar of Society: the Nature and Dynamics of Social Norms*. New York: Cambridge University Press

Bicchieri, C. and Mercier, H. (2012). "Norms and beliefs: The dynamics of change", in B. Edmonds (ed.) *The Dynamic View of Norms.* Cambridge: Cambridge University Press

Bicchieri, C. and Muldoon, R. (2011). *Social Norms*. The Stanford Encyclopedia of Philosophy

Coleman, J. (1990) *Foundations of Social Theory.* Harvard: Belknap.

Feldner, Y. (2000). "Honor Murders. Why the Perps Get off Easy," *Middle East Quarterly*, pp. 41-50

Fishbein, M. (1967) "A consideration of beliefs and their role in attitude measurement." In *Readings in attitude theory and measurement*, Fishbein, Martin, ed. New York: Wiley.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA, Cambridge University Press.

Miller, D.T., and McFarland, C. (1987). "Pluralistic ignorance: When similarity is interpreted as dissimilarity". *Journal of Personality and Social Psychology, 53,* 298-305.

Parsons, T. (1951) *The Social System*. New York: Routledge.

Sugden, R. (2001). "The Bond of Society: Reason or Sentiment? *Critical Review of International Social and Political Philosophy* 4 (4):149-170.

Wikan, U. (2005). "The Honor Culture", Karl-Olov Arnstberg and Phil Holmes, trans., originally published as *En Fraga Om Hedre* (A question of honor). Stockholm: Ordfront Forlag