politics, philosophy & economics article

© SAGE Publications
Los Angeles, London,
New Delhi, Singapore
and Washington DC

1470-594X
201008 9(3) 297–313

Norms, preferences, and conditional behavior

Cristina Bicchieri
University of Pennsylvania, USA

abstract This article addresses several issues raised by Nichols, Gintis, and Skyrms and Zollman in their comments on my book, *The Grammar of Society: The Nature and Dynamics of Social Norms*. In particular, I explore the relation between social and personal norms, what an adequate game-theoretic representation of norms should be, and what models of norms emergence should tell us about the formation of normative expectations.

keywords social norms, personal norms, expectations, experimental games, Bayesian games, evolutionary models

The articles contributed to this issue touch several crucial matters that need to be addressed by a theory of norms. The first is essentially an empirical question: how do we know, when we observe behavior in the laboratory, that participants are following a norm, and if so, is it a personal or a social norm? The second has to do with the game-theoretic representation of norms as equilibria, and the issue raised here is what sort of games best represent norms as opposed to, say, conventions. Finally, there is the question of how norms emerge, and how to model such a process convincingly. I have tackled some of these issues in my book and subsequent papers,[1] so in what follows I will discuss my most recent results, as well as suggest possible ways to proceed to address what are still very open questions.

The personal and the social

Facing experimental data that defy the hypothesis of self-regarding preferences, experimental economists have focused their attention on so-called social preferences, and even when admitting that social norms may indeed play a role in influencing behavior, nobody has gone beyond this generic assertion. This is not

DOI: 10.1177/1470594X10369276
Cristina Bicchieri, Philosophy Department, 491 Cohen Hall, University of Pennsylvania, 249 S. 36th Street, Philadelphia, PA 19104-6304, USA [email: cb36@sas.upenn.edu]

surprising, since to show that social norms play a role in decision-making, one needs to have an operational definition of what a 'norm' is. Providing such a definition is important, as it allows us to make testable predictions about the conditions under which a norm will be followed, besides giving us a way to determine that a norm is indeed present. My definition of social norm is based on conditional preferences and two kinds of expectations: empirical and normative. By empirical expectations I mean the belief that enough other people in a similar situation obey the norm (or have done so in the past). By normative expectations I mean the belief that enough other people think we ought to obey the norm in that situation, and may even be willing to sanction us in a positive or negative way depending on our choice to obey or contravene the norm. It is important to note that I say 'enough other people', and not a majority or even a large majority of people. This is because different individuals will have different thresholds below which they will consider the number of norm followers too small to count. Furthermore, even the same individual may have different thresholds for different norms. In a small gang, the threshold may be quite high, and so it may take a very small number of deviants to undo a norm of violence. For well-established, well-entrenched social norms, such as reciprocity norms, the threshold is usually lower, but again our personal allegiance to such norms will determine what a 'sufficiently large number of followers' means for each of us.

Since social norms typically pertain to situations in which there is an inherent conflict between individual and collective interests, such as those depicted in social-dilemma, trust, and ultimatum games, empirical expectations of others' compliance with the relevant norm may not be sufficient to induce conformity, and the presence of normative expectations may be in order. As I made clear in my definition of normative expectations,[2] it is possible that, for some individuals, the presence of sanctions is not necessary to induce them to conform. Such individuals will believe that others' normative expectations are legitimate, and will feel obliged to fulfill them. This belief, however, will not give them an independent reason to obey a norm once they come to realize it is no longer widespread. Indeed, having conditional preferences[3] implies that one may follow a norm in the presence of the relevant expectations, but disregard it in their absence.

My definition of norms lends itself to a series of testable predictions.

1. Since expectations matter to choice, *manipulating expectations* (via changes in information) will result in very different behavioral outcomes (this effect is not captured by inequity-aversion theories).
2. We will observe intra-subjective variation in behavior even if monetary consequences remain the same.
3. When a norm can have several different interpretations, there will be *self-serving biases*, that is, individuals will choose the interpretation that most favors them. For example, if the utility of abiding by a norm of equity is higher than the utility of following a norm of equality, we would expect an agent to choose the first.

4. Whenever it is possible to defy normative expectations without consequence, *norm-evading* behavior will be more frequent. For example, if agents can cheat without being found out, we would expect more cheating behavior.
5. There should be more variability of individual behavior in games in which it is not clear what the relevant norm is or in which there are competing norms.
6. Variability should decrease as more cues pointing to a particular norm are supplied. So there should be more variability in a dictator game, in which there may be no obvious norm, than in an ultimatum game.

In the experiments I performed with Erte Xiao and Alex Chavez, several such predictions were put to the test. In recent work by Bicchieri and Chavez (2010), participants were presented with three mini ultimatum games.[4] In each game, the proposer had three allocation choices: (5,5), (8,2), and the toss of a fair coin. In the coin case, heads meant (5,5) would be the outcome and tails instead resulted in (8,2). The experiment was double blind, participants were randomly matched and knew it, and it was also common knowledge that each game would be played with a different partner, so no learning was involved.[5] The only difference between the three games was the information the participants received. In the public information condition, it was common knowledge that the proposer had three choices, and if a proposer chose to toss a coin, both her coin choice and the final outcome were known to the responder. In the private information condition, the proposer had the same three choices, but the responder did not know it and the proposer was aware of this informational asymmetry. In this condition, the responder believed the proposer to have only the (5,5) and (8,2) choices, and even if the proposer could still (secretly) choose to toss a coin, he knew the responder would interpret the result of a coin toss as an intentional choice. In the limited information condition, it was again common knowledge that the proposer had three choices, but this time the informational asymmetry lay in the fact that the responder could not tell whether his offer was the result of a coin toss or an intentional choice on the part of the responder. The results are quite striking, as we observed the same players behave very differently when provided with different (symmetric or asymmetric) information. The question raised by Nichols is whether such behavioral changed can be explained by the conditional allegiance to a personal, as opposed to a social, norm of fairness.

Now suppose an individual has a personal (as opposed to social) norm of fairness. As I discussed in my book, having a personal norm would give one an independent reason to follow it, that is, a reason that is independent of one's expectations of others' compliance.[6] This, however, does not mean that one is completely indifferent to what others do or expect one to do. It just means that the personal-norm follower is more resistant to social influence than a social-norm follower. Even if one has independent reasons to act fairly, there might be circumstances in which these reasons are overridden by other, stronger motives. This may happen when we face conflicting norms and have to choose among them. For example, I may have a strong personal fairness norm, but be aware that social interactions in the culture to which I belong are governed by a norm of reciprocity. There are circumstances in which reciprocity would make me give someone a larger share than fairness dictates, and I may decide that (for the sake of social harmony) this is what I will do. Alternatively, it may be the case that I realize that my personal commitment to fairness is defeated by the fact that the people I am interacting with have a very different view of what an acceptable share is. In Calvino's novel *The Cloven Viscount*,[7] the good half of the viscount has strong moral views and is totally committed to fulfill his moral duties. His life is devoted to mending the dreadful deeds performed by his evil half. The sad truth is that his unequivocal commitment ultimately becomes a burden to the beneficiaries of his good will, who feel belittled and in a permanent state of obligation to repay his good deeds. Had the viscount realized what misery his actions were producing, he might have refrained from strict adherence to his unyielding personal norms. In both examples, it would seem reasonable to relax one's allegiance to a personal norm. Yet there is a crucial difference between conditional allegiance to social as opposed to personal norms: whereas a social-norm follower may be induced to cheat if her actions cannot be monitored and punished, I would expect a personal-norm follower to refrain from deception, since she would have good, independent reasons to behave correctly in this case.

According to Nichols, if a person is motivated by a personal preference for behaving in a fair way, such a preference would be active if no other, stronger alternative preference is at work.[8] In his example, one may value giving to charity, but if one is reasonably sure very few others give to charity, then one's preference for charitable giving may be superseded by another preference, for example spending the money in a fancy restaurant. In this case, recognizing that others do not give significantly dampens one's desire to be charitable. In the Bicchieri-Chavez experiment, however, there is no empirical expectation to draw upon: the only expectations that matter are normative ones, and they refer to what the responders believe is fair. In this case, a personal norm follower would have no reason to abandon fairness and start cheating.

Note that in order to differentiate between personal and social norms, we have to be able to tell that a social norm exists and is shared in the present circumstances. In the Bicchieri-Chavez experiment, this assessment is based on the existence of mutually consistent second-order beliefs. To be more specific, we did not just ask responders what sort of division *they* thought to be fair. The answer to such a question would tell us what their personal normative beliefs were, but we would be unable to assess whether a social norm was present. To gauge the presence of a social norm, it is important to ask people not just what their personal normative beliefs are, but what they expect *other's* normative beliefs to be. In the context of our experiment, Chavez and I wanted to know not just what the responders personally thought was fair, but also what both proposers and responders thought the responders believed to be fair. That is, it is the consistency of second-order beliefs that tells us that a norm is present. Interestingly, we found a remarkable consistency in the assessments of proposers and responders. Both groups believed that almost all responders would find (5,5) to be fair, but they also believed that a majority of responders would find the tossing of a coin a fair way to allocate the money.

So suppose, again, that one has a personal norm of fairness, and could 'cheat' the responder in the limited condition by offering only $2, hoping that the responder will accept it, believing that it is the result of a coin toss. In this case, there is no information as to what other proposers will do, but if we look at the proposers' beliefs about what responders consider fair, we see that most responders are (correctly) believed to think that tossing a coin is a fair option. In this case, if one expects $2 to be accepted by a responder who believes (or wants to believe) that it is the result of a coin toss, one may offer $2 with impunity. Indeed, 76 percent of the proposers choose the (8,2) division in the limited condition, whereas only 44 percent choose (8,2) in the private condition and just 32 percent choose it in the public information condition.[9] It should also be noticed that a very common offer pattern is the choice of (5,5) in the private information condition, tossing a coin or (5,5) in the public condition, and (8,2) in the limited information condition. Since the experiment involves the *same* players playing three ultimatum games with different partners, such behavioral changes suggest that the information manipulation allows many participants to skirt a fairness norm when they can do it with impunity. Could such players be said to have a personal (as opposed to a social) norm of fairness? I believe not, since, as I shall explain, in this case *there would be no reason to deviate from one's personal norm.*

To make my point clear, let us assume, with Nichols, that we are indeed conditional followers of our own personal norms of fairness. That is, we may behave fairly under certain conditions and unfairly in others. When would we expect a personal norm of fairness to be overridden by other, more selfish considerations?

1. When a majority of people are behaving unfairly. In this case, one may feel that one's efforts to be fair are useless, as one's fair action will bring no change and probably most people do not even expect to be treated fairly. So, for example, if I am told, prior to making my proposal in an ultimatum game, that a large majority of proposers just offer $3, I may follow their example, since I may infer from this information that most responders accept $3. In this case, it may not be possible to distinguish between those who follow a social versus a personal norm. Both are conditional on expecting others to behave in a fair way.
2. When a majority of people do not believe one should act in a fair manner. In this case, others' normative expectations are absent. Suppose, again, that I believe one ought to behave fairly, but I find myself in a situation in which, because of cultural reasons, the majority of ultimatum game proposers believe that those who have the power to decide how to divide the money should keep

> more than half, indeed they should keep 70 percent of the money, because they are endowed with this special power. Again, in this case I may infer from the normative information that responders are accustomed to getting $3, and thus will not get upset and accept it. In this case, too, social and personal norms cannot be differentiated.

In the two situations just described, since individuals have only limited information, empirical or normative, they may not differentiate between empirical and normative expectations. Indeed, in the presence of a single normative piece of information (that is, what people think is appropriate behavior), we *have no reason* to believe that people will act in ways that conflict with what they say they should do. Similarly, if the only information we have is actual behavior, we have no reason to think that the behavior we observe is at odds with what the actors think is appropriate behavior. If there is no other relevant information, the simplest assumption to make is that behavior and the beliefs that support it are correlated. Usually, people infer normative expectations (empirical expectations) from empirical information (normative information) when that is the only information available. This explains the correlation Bicchieri and Xiao found in the 'single information presentation' treatments.[10] This also motivated the need to present subjects with conflicting normative and empirical information, the idea upon which our experiment is founded. Through a study of behavior in the 'conflicted' conditions one can tease out any differential behavioral impacts of different types of expectations.[11]

It is an entirely empirical question how much an individual who follows a personal, as opposed to a social, norm would be swayed by being presented with conflicting information. However, I want to advance the hypothesis that whenever the normative expectations of others are explicit and in line with those of the individual in question, the presence of transgressive behavior (the conflicting empirical expectations) should have less of an impact on her choices. Though in this case a social-norm follower will find an excuse to evade the norm, I would expect the holder of a personal fairness norm to be less prone to being influenced. Specifically, I would expect personal-norm followers to display a *higher percentage* of fair choices in the conflicting selfish choice and fair belief (SC + FB) condition than when facing the selfish choice (SC) condition alone, and this percentage should be close to the percentage of fair choices in the fair belief (FB) condition. The social-norm follower, on the contrary, should display very similar behavior in SC + FB and SC, since the presence of a majority of selfish choices would give her an excuse to override the pull of normative expectations of fairness.

The most convincing evidence about the large presence of social-norms followers, however, comes from the experiment discussed in Bicchieri and Chavez.[12] In this case, someone who has a personal norm of fairness should not change her behavior depending on the information condition, and in particular should not cheat the responder in the limited condition. From the data it appears that most participants follow what I call a social norm, and are prepared to shirk it if they can do so without incurring a cost. By giving an equal share (or tossing a coin) in the private and public conditions, these subjects show they are aware of the normative expectations of the responders, so in this case there is no available excuse regarding the presumed absence of such expectations.

With these comments I do not wish to deny the presence of personal norms, far from it. I am prepared to agree with Nichols that our following a personal norm may be conditional upon having specific expectations regarding others' compliance or expectation of compliance. Yet, as I have suggested, there are limits to how sensitive a personal-norm holder is to empirical and normative expectations. In our experiments, the large majority of participants are very responsive to information manipulation, but there is always a small percentage of outliers, in both directions. These people do not seem to care about what others do or expect them to do. Among those who are sensitive to empirical and normative expectations, the majority are prepared to cheat if given the opportunity. There is no evidence that these individuals hold a personal norm of fairness.

The issue of how a norm becomes moralized, or as I like to say, has become deeply entrenched not only in a culture, but also in the psyche of its members, is another interesting question raised by Nichols that merits serious consideration. One way in which moralization can be explained is by reference to the significance of the social function that a particular norm plays. Another way refers instead to the emotional resonance that some norms seem to have. Certain prescribed or proscribed actions, for example, are independently linked to some common, natural emotions, such as disgust or empathy. In the emotivist view, these emotions contribute to the cultural success, resilience, and moralization of certain norms.

According to the first strand of reasoning, moralized norms are linked to some socially valuable function they perform. For example, the norms of etiquette studied by Norbert Elias or the foot-binding norm prevalent in China for centuries came to have an important signaling function.[13] Table manners came to signal inclusion in the ruling class, and foot-binding signaled regard for custom and tradition, an essential Confucian value. Violating the norm amounted to an infringement of the social order, the transgressive act being perceived as the tangible sign of a character flaw. Another interesting example is the presence, in many past and present societies, of equity norms, such as the requirement of giving a larger share to the deserving or to the needy. When there is no reason or possibility to differentiate among claimants, fairness usually means an equal allocation among the parties. When, instead, it is possible and reasonable to discriminate among claimants according to some relevant characteristic, need and desert are commonly chosen. Social groups may offer a different interpretation of *who* is needy or *what* merit or desert mean, but it is almost universally recognized that some criterion of need or desert (or both) has to be adopted if the circumstances warrant it. Though the existence of a single fairness rule, such as equality, would guarantee easily reachable agreements on the terms of any distribution, such parsimony would come at a huge social cost. Those who are in greater need (because of conditions, physical or otherwise, out of their control) would receive much less than required to sustain them, and the lack of a system of social insurance (guaranteeing a minimum endowment to the less well off) would almost certainly generate widespread social instability. On the other hand, societies need to provide incentives for their most able and productive members to perform in such a way as to increase the collective well-being. Equity norms thus fulfill important social functions and often become moralized. The deserving person who is denied his due feels rightfully angry; indifference to the plight of the needy evokes contempt, and guilt in those who decline to help or share.

Providing a social account of how moralization of a norm may develop is different from explaining it as the result of basic emotional dispositions. The evidence about the involvement of affective areas of the brain in norm-related behavior[14] would seem to make a strong case in favor of an emotivist explanation not just of our motivation to comply with norms, but also of the resilience, entrenchment, and moralization of norms such as equal division. Yet characterizing brain areas as ‘affective’, as opposed to ‘cognitive’, is misleading, as there is no clear-cut distinction between such areas, and each area of the brain is typically involved in many tasks. For example, the anterior insula is activated in disgust, but also in disbelief,[15] and the amygdala is not just involved in fear, but also in visual attention, a task that can hardly be described as affective. Moreover, even if some brain areas are more involved than others in affective responses, judgment and decision tasks simultaneously activate several areas of the brain, both affective and cognitive ones. Spitzer et al. have shown that, when third-party punishment is introduced, participants in a dictator game tend to behave more fairly, and their behavior is associated with a stronger activation of the dorsolateral prefrontal cortex (DLPFC), which is involved in cognitive control.[16] In fact, this area of the brain is active in all decision-making tasks, and has been shown to be involved in norm compliance. In ultimatum games, neural activation involves the anterior insula (commonly associated with anger), the DLPFC, and the anterior cingulate cortex (involved in reward anticipation, error detection, and modulation of emotional responses), leading to the conclusion that the split between cognitive and affective is rather artificial, as all these areas are jointly involved in processing the information that ultimately leads us to make a fair offer or reject an unfair one.

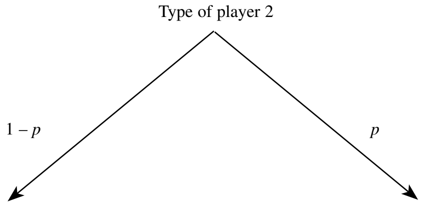
The decision to reject an offer, for example, implicates a belief that a fairness norm should apply, a negative evaluation of the offer, the violation of an expectation considered legitimate, and an emotional response to the violation. As I discussed in my book, when we encounter a new situation, such as an experimental game, we have to interpret, understand, and encode it.[17] This takes a series of steps, from categorization to the elicitation of scripts or schemata. Categorization activates a comparison process to assess the similarity of the present situation with members of a category stored in memory. What enters in our categories is determined by the culture and society we live in, so it is not surprising that in different cultures, and particularly in small societies that significantly differ from ours,[18] we witness very different interpretations of simple games such as the ultimatum and trust games. If the Au and Nau of New Guinea make 'hyperfair' offers that are rejected, it is because they categorize the ultimatum game along with other exemplars of gift giving. The gift giver enhances his status within the community (so the larger the gift, the higher the status), whereas the recipient incurs an obligation to eventually reciprocate. Once a situation is categorized as a member of a certain category, a schema (or a script) is invoked. A schema is a cognitive structure that represents knowledge about people, events, and so on. It involves beliefs, expectations, and even behavioral rules. A script is a schema for a social event: it describes a stylized, stereotyped sequence of actions and defines actors and roles. Our scripts allow us to make inferences about unobservable variables, predict behavior, make causal attributions, and modulate emotional reactions. Scripts are a source of projectable regularities as well as the legitimacy of our expectations.

Social norms are embedded into scripts that define (among other things) expectations and the kind of emotional reactions that are most appropriate to expected and unexpected occurrences. Thinking in terms of scripts leads us to reject a sharp distinction between affect and cognition, as both are intertwined in our stereotyped mental models of what fair divisions are. Indeed, scripts provide a bridge between social function and emotional response, and do so in such a way that does not privilege either one. Identifying the neural mechanisms involved in norm compliance and evaluation of norm transgression helps us understand how we process information and arrive at judgments and decisions, but gives us no reason to throw away an explanation in terms of preferences and beliefs in favor of one that takes emotions as primitive.

Games and evolving norms

Game theory has been a useful tool to model various aspects of social norms. Once a norm is present, its continuous grip on us will depend upon a web of expectations, and as long as such expectations are self-fulfilling, we will keep conforming to the norm. A norm can be thus represented as an equilibrium, in the sense that each player maximizes her expected utility if she takes the actions of the other players as given, and the players' beliefs are correct.[19] Which type of game best depicts this situation? In *The Grammar of Society*, I said that the existence of social norms, and especially the pro-social norms I am interested in, *transforms* mixed-motive games.[20] The latter games illustrate the ever-present conflict between individual and collective interests, a conflict that is usually mitigated, if not solved, by the presence of pro-social norms. The existence of a norm makes it possible, depending upon one's empirical and normative expectations, to get outcomes that are better for all the parties involved than the outcomes that are possible without the norm. Take as an example a very simple, one-shot trust game in which the first mover (trustor) is endowed with \$10 and has only two choices: to trust and thus give the \$10 to an anonymous trustee or not to trust and thus keep the money. If she trusts, the money is multiplied by a factor of three. In that case, the second mover (trustee) receives \$30 and faces two possible choices: to keep all the money or to give back half to the trustor. Clearly, trusting and reciprocating would make both better off, as each would end up with \$15. However, if we make the usual assumption of self-regarding preferences, rationality, and common knowledge thereof, we can predict that no trust will occur, as reciprocation could not be expected.

Now suppose there exists a reciprocity norm and the players are aware of it. In this case, *if* the trustor expects the trustee to be a norm follower, then it would make sense for him to send the \$10. If a norm of reciprocity exists, what the trustor now faces is a Bayesian game in which, with probability $1 - p$, her counterpart is not going to follow a reciprocity norm.[21] We may say that, in this case, the trustee is the selfish type of player and the trustor will still be playing the usual trust game. However, with probability p the trustee follows a reciprocity norm, and is thus trustworthy. In this case, the trustor would do better trusting, as a trustee who is a norm follower would get a lower payoff by failing to reciprocate (NN). The situation is depicted in Figure 1. The trust game has a unique Nash equilibrium: (NT, NR). In this case, the players are trapped in an inferior outcome. The new game generated by the existence of a reciprocity norm has two Nash equilibria, (T, N) and (NT, NN), but now (T, N) is better for both players. The second mover, if he is a norm follower, will prefer to comply with the norm when it applies, as depicted by the lower utility he gets by violating the trust of the first mover (NN).[22] If the trustor assesses a probability greater than .67 that the trustee is a norm follower, then it makes sense to trust.[23]



	R	NR
T	15,15	0,30
NT	10,0	**10,0**

Trust game

	N	NN
T	**15,15**	0,5
NT	10,0	**10,0**

Norm Game

Figure 1

We have significant experimental evidence that a majority of first movers (trustors) in trust games do trust, even if trust is not fully reciprocated.[24] We also have evidence that a lack of reciprocation in trust games elicits third-party punishment and that there is an overwhelming agreement among third parties that lack of reciprocity *should* be punished,[25] which tells us that reciprocity is a social norm. Bicchieri, Xiao and Muldoon, in a recent article, advance the hypothesis that trusting acts as a signal that is intended to focus the recipient on a reciprocity norm.[26] If such a norm exists and is shared, then it is rational to trust insofar as one believes that in so doing one will trigger reciprocation even when the material incentives to reciprocate are absent. Yet my analysis of the conditions under which a norm will be followed also explains why less than a majority of trustees reciprocate, in that the absence of punishment (the game is one shot and anonymous) would induce some players to skirt the norm. Still, a significant proportion of players *do* reciprocate, which means that a proportion of the population obeys a reciprocity norm even in the absence of obvious sanctions.

Gintis has proposed a different model, one according to which a trust game or an ultimatum game would be part of a larger game that has many more strategies than the original one.[27] Agents correlate their strategies to an event that is external to the game and, provided that they have common priors and common knowledge of the game, their rationality and their beliefs, a correlated equilibrium obtains.[28] That is, conditional on accepting the recommended strategy, an agent's expected payoff for playing that strategy is no worse than the payoffs obtained by playing any other strategy. Such an equilibrium would be better than the Nash equilibria of the original game; for example, the equilibrium in which all players follow a social norm of reciprocity is certainly collectively more advantageous than the original (NT, NR) equilibrium. In this context, social norms should be understood as the 'choreographers' of such a new equilibrium, not as the selection devices of simple Nash equilibria. Correlated equilibria have been proposed as models for conventions,[29] and this is a perfectly adequate model for situations in which agents jointly decide to use some external signals to coordinate their strategies. Conventions, however, can also emerge out of repeated interactions in which players eventually converge on a particular strategy profile. In this case, we may still think of them as Nash equilibria of original coordination games.

In any convention, it is rational for an agent to coordinate with what he expects others to do and, as I have discussed in my book,[30] the empirical expectation that others follow the convention is all that one needs in order to conform. But would a social norm telling players to reciprocate trust be so effortlessly followed? Social norms are different from conventions, in that obeying a norm usually has a cost, and normative expectations are a crucial part of the picture. If we are starting with mixed-motive games, and there is reason to believe that we are in a noncooperative context (as when interactions are anonymous), the presence of normative expectations helps understand how agents may be motivated to behave in a pro-social manner.

Gintis's model cannot apply to social norms unless, as he must assume, players have other-regarding preferences, or a 'normative predisposition' to obey social norms. This assumption, however, is not supported by experimental evidence. I do not deny that individuals have evolved the capability to learn and apply social norms even to situations that are completely new, but there is much evidence that we are *conditional followers* of norms. In fact, as the experiments reported in Bicchieri and Xiao (2009) and Bicchieri and Chavez (2010) show, manipulating information produces major changes in behavior, and the existence of a social norm is no guarantee that it will be followed. The real challenge we face is to explain how normative expectations emerge or, in other words, how the beliefs that support social norms take shape.

The article by Skyrms and Zollman offers interesting suggestions about the directions evolutionary analysis should take.[31] Even if they do not explicitly mention them, they recognize the importance of categories and scripts, insofar as framing 'should be interpreted as a signal about the relevant class of social interactions'. Evolutionary analysis, in their view, should be directed to systems or classes of social interactions. Given the existence of classes of social interactions, however, are we justified in claiming, as they do, that individual behavior across strategic contexts is usually very similar, and thus conclude that norms are often insensitive to context? Quite to the contrary, we know that context, and in particular the way the experimenter describes the game, matters: if we describe a social-dilemma game as a 'community game', the rate of cooperation is very different than if we describe it as a 'market game'.

I believe that Skyrms and Zollman's statement about similar behavior across contexts needs to be qualified in order to be correct. Thinking in terms of classes of social interactions may induce us to think that what has evolved are 'generic' norms that apply across a wide variety of strategic contexts. It would surely seem more economical to develop strategies that are not game contingent, given the cognitive cost of differentiating among different game forms. This idea, however, flies in the face of much experimental evidence, and seems to contradict Skyrms and Zollman's highlighting the importance of framing effects. If a single, generic norm of fairness is what has evolved, then people should be indifferent to the obvious cues that differentiate a dictator from an ultimatum game. But are they?

Even if in our culture both the Nash bargaining game and the ultimatum game tend to converge toward a 50-50 split in the laboratory setting, the dictator game is an entirely different story. In the first two games, the threat of punishment looms large, whereas in the dictator game the second mover is a passive recipient of whatever the proposer decides to offer. Very often proposers in dictator games offer nothing, and we rarely see offers greater than 20 percent. The only cases in which we observe fairer offers are those in which the set of proposer's choices is restricted to a fair offer (that is, (5,5)) and a very unfair offer (such as (8,2)). In this case, the experimenter has cued a fairness norm, and the entire context has changed from one in which there is basically no rule to one in which there is a clear-cut choice between being fair or unfair.

The claim of Skyrms and Zollman's article about the existence of classes of social interactions might be interpreted as referring to social categorizations. We seem to possess encompassing social categories that include several types of interaction, so an ultimatum game played in the laboratory will be seen as relevantly similar to a class of situations in which we have to divide something and, if there is no sufficient information to deviate from an equal-division rule, that is what we adopt, especially since we expect retaliation if we deviate from it. Yet when we add 'cues' as irrelevant as stars[32] or assign roles (proposer and responder) on the basis of a trivia quiz, a different interpretation of fairness seems to apply: fairness as equity (that is, the more deserving should get more) overrules the simpler equal-division rule. The same shift to equity interpreted as 'give more to the needy' can be seen in games in which the players are aware that one of them is in need. Just because we use the word 'fairness' in all these contexts does not mean we are referring to the same rule in all of them. As I discussed in my book, the same category that encompasses 'fair divisions' may activate different scripts, depending on the cues provided by the situation at hand.[33] If a fair-division category may trigger several scripts, this would explain our fine-tuned responses to the description of a situation. Skyrms and Zollman correctly interpret cues as signals that a script applies to a particular situation. What remains to be explained is why we have settled on a specific partitioning of the social world and why we have developed, in the case of fairness, several different rules of fair division.[34] There is no unique, generic norm of fairness, just a family of fair-division rules that dictate very different behaviors across different contexts.

It is important to notice that fairness norms that concern merit or need involve another kind of *signaling*. That is, one has to signal need or merit in order to obtain a differential share, but such signals are far from perfect. Indeed, there is the ever-present possibility of cheating. The further we move away from simple physical prowess or obvious conditions of need, such as infancy or old age, the more 'noisy' the signals become. Bicchieri and Muldoon have argued that when norms involve signaling, multiple competing norms in a population can be socially optimal and dynamically stable.[35] Our argument is based on the fact that competing norms lower the cost of enforcing honest signaling. We do not, however, have a story about how such a family of norms emerge, and it is a story that needs to be told.

A significant new approach to norm emergence highlights the importance of structured interactions.[36] What is investigated in these models is the effect of specific social structures on the equilibrium outcomes of many different games. This is an important step away from random matching, the usual assumption in many evolutionary games, and it has the advantage of showing, for example, how certain norms can emerge in particular network structures, but not others. These models, however, tend to identify a norm with a strategy. Though this is certainly the case with fairness norms, other norms, such as reciprocity, seem to involve a family of strategies. When we observe a population that consistently trusts or reciprocates, it would be a mistake to assume that what is at work is a norm adopted by the entire population. As Bicchieri, Duffy and Tolle (2004) have demonstrated, such a uniform behavioral pattern is supported by a family of conditional strategies.[37] Each of these strategies punishes lack of reciprocity to different degrees, but it is their combination that allows the pattern to emerge and survive. In this case, as in the case of other cooperative norms, it would be a mistake to identify a norm with a single strategy. Still, such an evolutionary account is limited in that it explains how certain *behavioral regularities* have evolved, but does not help in understanding how the psychological and cognitive aspects of norm following evolve. In particular, no one has a plausible story for how normative expectations come to be. All the models, however, implicitly assume that people have some version of such expectations in the background, to be modified via learning.

Another important piece of the emergence puzzle is thus an explanation of how individuals come to recognize that a norm exists, and how they come to form the kind of expectations that support norm compliance. In the last chapter of my book, I report the results of a simulation I ran with Jason Alexander.[38] We focused on a norm of fair division, and crucially assumed that individuals playing repeated ultimatum games share certain basic psychological dispositions, such as herding behavior, that individuals normally display in new or uncertain situations. Thus, we assumed that the interacting players would look for behavioral regularities even if, at least at the outset of their interactions, none existed. Each player faced with a monetary-division task will try to discover what behavioral regularity, or 'norm', exists before acting, since by following the norm he can legitimately expect his offer to be accepted. Once a player believes he has identified a norm, he will tend to follow it, provided he believes a sufficient proportion of the population likewise follow it, and also believes that this proportion of the population expect him to conform to the norm. In the ultimatum game we have studied, not following the norm may mean that one's offer is rejected, and hence one receives a payoff of zero. Our account of norm emergence is more complex than the usual models, in that we used a norm-based utility function[39] and made assumptions about individuals' different sensitivities to the 'putative' norm, as well as about the presence of different individual thresholds for coming to decide that a norm exists. It is remarkable how more realistic psychological assumptions, combined with population heterogeneity (with respect to sensitivity and thresholds), lead best-response players to converge toward an equal-division rule.

The heterogeneity of agents' sensitivity to the social norm, and in particular the distribution of this parameter, is what drives compliance with a social norm in the model I just described. What remains to be explained is how such a distribution of psychological types has evolved. It is likely that our sensitivity to social norms and the accompanying disposition to punish transgressors have evolved out of social-dilemma-type situations. Every social group, in order to survive, must produce and preserve a certain number of public goods. No such goods would ever be produced and sustained without a collective cooperative effort. The specific forms of cooperation may vary across time and cultures, but the propensities to identify and conform to norms as well as to be prepared to punish defectors are evolutionarily necessary to the very existence of public goods and society itself (the ultimate public good). We can live without fairness, but we cannot survive without cooperation.

I hope I have made clear that I do not consider the propensity to follow social norms an unconditional one. We observe too much opportunistic behavior and norm evasion to believe that we have evolved such a propensity. On the contrary, our compliance with norms is conditional on having certain empirical and normative expectations. Any theory of norm formation is incomplete without an explanation of how such expectations (the normative ones in particular) evolve.

notes

I wish to thank Ryan Muldoon, Erte Xiao, and Giacomo Sillari for many useful comments and suggestions.

1. C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge: Cambridge University Press, 2006); C. Bicchieri and A. Chavez, 'Behaving as Expected: Public Information and Fairness Norms', *Journal of Behavioral Decision Making* 2 (2010): 161–78; C. Bicchieri and R. Muldoon, 'Competing for Fairness', mimeo (Philadelphia: University of Pennsylvania, 2010); C. Bicchieri and E. Xiao, 'Do the Right Thing: But Only If Others Do So', *Journal of Behavioral Decision Making* 22 (2009): 191–208; C. Bicchieri, E. Xiao and R. Muldoon, 'Trust If You Wish, Always Reciprocate', *Politics, Philosophy and Economics* (forthcoming).
2. Bicchieri, *The Grammar of Society*, p. 11.
3. Having conditional preferences implies having a plurality of rank orderings. However, a challenge faced by any such theory is to model how preferences are indexed to classes of situations. For example, if my preference for following a fairness norm depends upon having certain beliefs, I may prefer x to y in a class of situations in which those beliefs are met, and y to x otherwise. Modeling an agent's different preference orderings is complicated by the fact that, due to threshold dependency, we are dealing with discontinuous functions.
4. See Bicchieri and Chavez, 'Behaving as Expected'.
5. It was also common knowledge that proposers, at the end of each game, would not be told the outcome of the game they played. Finally, players were paid for two out of three games, and the games to be paid for were randomly decided at the end of the experimental session.
6. Bicchieri, *The Grammar of Society*, p. 21.
7. Italo Calvino, (1951). *The Nonexistent Knight and the Cloven Viscount*. Torino: Giulio Einaudi Editore.

8. S. Nichols, ‘Emotions, Norms, and the Genealogy of Fairness’ (in this issue).
9. The trend is the same in the salient condition, in which we ask participants questions about what they think the responders believe is fair. In this treatment, there are more (5,5) and less (8,2) choices, but still the large majority of (8,2) choices is concentrated in the limited information condition.
10. Bicchieri and Xiao, ‘Do the Right Thing’.
11. In our experiment, we gave different information to different groups of subjects before they made an allocation choice in a dictator game. A group was told that a majority of subjects in a previous experiment made a fair choice (FC); another group was told that a majority of subjects in a previous experiment made a selfish choice (SC); a third group was told that a majority of subjects in a previous experiment believed that one should make a fair choice (FB); a fourth group was told that a majority of subjects in a previous experiment believed that one should make a selfish choice (SB); and the last two groups were given conflicting information. One group was told that a majority of subjects in a previous experiment believed that one should make a fair choice, but a majority of subjects in another experiment made a selfish choice (SC + FB). The last group was told that a majority of subjects in a previous experiment believed that one should make a selfish choice, but a majority of subjects in another experiment made a fair choice (FC + SB).
12. Bicchieri and Chavez, ‘Behaving as Expected’.
13. N. Elias, *The History of Manners* (New York: Pantheon, 1978); G. Mackie, ‘Ending Footbinding and Infibulation: A Convention Account’, *American Sociological Review* 6 (1996): 999–1017.
14. A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom and J.D. Cohen, ‘The Neural Basis of Economic Decision-making in the Ultimatum Game’, *Science* 300 (2003): 755–8.
15. S. Harris, S. Sheth and M.S. Cohen, ‘Functional Neuroimaging of Belief, Disbelief, and Uncertainty’, *Annals of Neurology* 63 (2008): 141–7.
16. M. Spitzer, U. Fischbacher, B. Herrnberger, G. Gron and E. Fehr, ‘The Neural Signature of Social Norm Compliance’, *Neuron* 56 (2007): 185–96.
17. Bicchieri, *The Grammar of Society*, p. 94.
18. J. Henrich et al. (editors), *Foundations of Human Sociality: Ethnography and Experiments in 15 Small-scale Societies* (Oxford: Oxford University Press, 2004).
19. A norm can also be instantiated by a set of strategies that are phenotypically similar. So there can be a one-to-many relationship between norms and equilibria. See C. Bicchieri, J. Duffy and G. Tolle, ‘Trust Among Strangers’, *Philosophy of Science* 71 (2004): 1–34.
20. Bicchieri, *The Grammar of Society*, p. 26.
21. In a typical Bayesian game, nature picks player types with a given probability, and players are assumed to share such ‘common priors’. See J. Harsanyi, ‘Games with Incomplete Information Played by “Bayesian” Players’, Parts I–III, *Management Science* 14 (1967–68): 159–82, 320–34, 486–502. In our case, we may think of shared scripts providing common priors.
22. The utility function I introduced tells us that, for a sufficiently high *k* (a parameter that measures one’s norm sensitivity), the trustee will prefer to reciprocate. See Bicchieri, *The Grammar of Society*, p. 52.
23. Note that the probability value .67 can be understood as the ‘risk factor’ of the

(T, N) equilibrium. Clearly, such an equilibrium is more risky than the (NT, NN) equilibrium.

24. The aggregate results show that the large majority of first movers transfer some money, but only close to 50 percent of second movers reciprocate (by giving back at least the amount sent). For example, Berg et al. found that 30 out of 32 first movers transferred some money, whereas only 14 out of 30 second movers gave back at least the transfer amount. See J. Berg, J. Dickhaut and K. McCabe, 'Trust, Reciprocity and Social History', *Games and Economic Behavior* 10 (1995): 122–42.
25. Bicchieri et al., 'Trust If You Wish, Always Reciprocate'.
26. Bicchieri, Xiao and Muldoon (2009) 'Trust If You Wish, Always Reciprocate'
27. H. Gintis, 'Social Norms as Choreography' (in this issue).
28. The common priors and common knowledge assumptions are rather severe constraints imposed on solution concepts. My model is much more general in terms of information conditions. In particular, my definition of social norms has the advantage of not assuming common knowledge on the part of the players. A norm only requires players to have mutual expectations (empirical and normative), and no assumption is made about how such expectations are formed or justified.
29. P. Vanderschraaf, 'Convention as Correlated Equilibrium', *Erkenntnis* 42 (1995): 65–87; P. Vanderschraaf, 'Knowledge, Equilibrium and Convention', *Erkenntnis* 49 (1998): 337–69.
30. Bicchieri, *The Grammar of Society*, p. 38.
31. B. Skyrms and K.J.S. Zollman, 'Evolutionary Considerations in the Framing of Social Norms' (in this issue).
32. S.B. Ball and C. Eckel, 'The Economic Value of Status', *Journal of Socio-Economics* 4 (1998): 495–514.
33. Bicchieri, *The Grammar of Society*, pp. 81 ff.
34. It may also be the case that different people may have different partitions, so even the assumption of common partitions is questionable.
35. Bicchieri and Muldoon, 'Competing for Fairness'.
36. B. Skyrms, *The Stag Hunt and the Evolution of Social Structure* (Cambridge: Cambridge University Press, 2004); J.M. Alexander, *The Structural Evolution of Morality* (Cambridge: Cambridge University Press, 2007).
37. Bicchieri et al., 'Trust Among Strangers'.
38. Bicchieri, *The Grammar of Society*, Ch. 6.
39. Ibid., p. 52.

Cristina Bicchieri is the Carol and Michael Lowenstein Professor of Philosophy and Legal Studies at the University of Pennsylvania. She works on judgment and decision-making, with special interest in decisions about fairness, trust, and cooperation as well as how expectations affect behavior.