Fairness and Social Norms

Daniel M. Hausman†

This essay comments on the theory of social norms developed by Cristina Bicchieri in *The Grammar of Society* (2006). It applauds her theory of norms but argues that it cannot account for the experimental results concerning ultimatum games. A theory of fairness is also needed. It develops a number of specific criticisms of her way of incorporating the influence of norms into preferences.

1. Introduction. In *The Grammar of Society* (2006), Cristina Bicchieri offers a theory of social norms that purports to explain many results of experiments by economists and psychologists and that aims to improve theories of economic behavior. She defends a view of social norms as constituted by the beliefs of individuals and conceptualizes the role of norms in influencing behavior by incorporating the influence of norms into individual utility functions that—as in orthodox economic models—represent individual preferences. Although Bicchieri is a philosopher, *The Grammar of Society* consists of substantial social-psychological theorizing.

This essay first distinguishes Bicchieri's theory of social norms from her proposal for incorporating the influence of norms into individual utility functions. I find her account of norms to be an impressive achievement, although I argue in Section 4 that it does not capture concerns about fairness. Qualms about her proposal for incorporating the influence of norms into individual preferences are developed in Section 3. Although I believe that Bicchieri's account has significant flaws, its theory of norms is intriguing and provides ingredients that may prove useful in future theorizing.

2. Bicchieri's Theory of Social Norms. Bicchieri's main contribution is a theory of social norms and an account of how they influence choice. What she calls "a social norm" matches some of everyday language and

†To contact the author, please write to: Department of Philosophy, University of Wisconsin–Madison, 5197 Helen C. White Hall, 600 N. Park Street, Madison, WI 53706-1474; e-mail: dhausman@wisc.edu.

Philosophy of Science, 75 (December 2008) pp. 850–860. 0031-8248/2008/7505-0030$10.00
Copyright 2008 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

social scientific usage, but she is not trying to capture what social scientists or ordinary people mean. According to her theory, social norms apply exclusively to circumstances in which there are conflicts of interest, some strategy combinations are collectively better than others, and some Pareto optimal strategy combinations are not Nash equilibria. When people follow social norms, they are motivated by their belief that others expect them to act in accordance with the norm. Without that belief, many would not find it in their interest to act as the norm requires. This fact distinguishes social norms from conventions and what Bicchieri calls "descriptive norms," in which it is in one's individual interest to conform, given one's beliefs about how others will behave (31ff.).

A social norm N governing situation S exists in a population P when sufficiently many members of P know that N applies in S and prefer to conform to N if (a) they believe that enough others conform and (b) they believe that enough others expect them to conform (and possibly, in addition, prefer them to conform and may sanction transgression). The norm is followed if conditions a and b (which Bicchieri calls, respectively, the empirical and normative expectations) are satisfied and, as a result, sufficiently many prefer to conform to N (11ff.). My notation here and elsewhere does not perfectly match Bicchieri's. Unanimity is not required, and members of P might disagree about how large a portion of P must hold the requisite beliefs and preferences. Bicchieri argues that the empirical and normative expectations of individuals motivate them to act in accordance with the norm not only because people dislike the disapproval of others and seek to avoid the sanctions others may apply but, more important, because people recognize the legitimacy of the expectations of others (21–25). Although embracing and following a social norm may be a deliberate act, often it is relatively automatic (48–51).

Unlike some sociological theories of compliance with social norms, an explanation of behavior that invokes social norms as defined by Bicchieri does not compete with an explanation in terms of preferences and expectations. On the contrary, an explanation in terms of social norms *is* an explanation in terms of the empirical and normative expectations of individuals and in terms of their preferences to meet the expectations of others.

If subjects in a laboratory for experimental economics follow a norm calling for maximizing group benefits, then it may be individually rational for them to cooperate in what looks like a prisoner's dilemma. To implement this explanatory strategy Bicchieri must address the following three questions:

1. How do experimental subjects come to believe that particular norms apply to the situations they encounter, and how do they form empirical and normative expectations?

https://doi.org/10.1086/594529 Published online by Cambridge University Press

2. How does the motivation to conform to social norms enter into individual preferences?
3. How can claims about the motivations and the empirical and normative expectations of experimental subjects be tested?

An underlying worry is that invoking social norms will trivialize one's explanations. Is not one in danger of being able effortlessly to accommodate anything? When an individual defects in a prisoner's dilemma, she is rationally serving her material interests, which motivate her more strongly than norm following. When another cooperates, he is rationally following a norm, which motivates him more strongly than his material interests. If neither behavior can be precluded, it is an illusion to suppose that one has explained anything. Bicchieri seeks to allay these worries with detailed discussions of testable implications of her account that conflict with the implications of alternatives. But some qualms remain.

Bicchieri's account of how experimental subjects come to believe that particular norms apply and how they form their empirical and normative expectations is (apart from some simulations in Chapters 5 and 6) informal. From the perspective of someone who is not trained as a social psychologist, her views in Chapter 2 seem very plausible: detailed cues concerning experimental situations lead individuals to assimilate them to some script or schema, whose instantiation generates the expectations that constitute social norms. The categorization of an experimental setup need not be obvious or uncontroversial, and one reason why communication among experimental subjects often calls forth behavior that is not materially self-interested is that it permits subjects to agree on a particular interpretation of a situation and on which norms govern. Because norms are specific and people are aware of a great many, details of the way in which a problem in presented to experimental subjects may provoke highly variable behavior.

With respect to the second problem of incorporating social norms into people's preferences, Bicchieri proposes a formal model (52–53) that is fully compatible with her view of social norms but to a considerable extent independent of it. All it requires of norms is that where they exist, they specify a strategy for each player in any given situation (where situations are defined by games and by the strategies followed by the other players). A strategy profile renders a norm applicable to an individual j (in Bicchieri's terms, "instantiates" a norm for j) if there is a norm specifying what j should do in this game, given the strategies the others actually follow. A strategy profile violates a norm if for at least one player j there is a norm applicable to j that j doesn't obey. Bicchieri's norm-included

https://doi.org/10.1086/594529 Published online by Cambridge University Press

utility function is then as follows:

$$U_i(s) = \pi_i(s) - k_i\pi_{\mathrm{ml}}(s).$$

I have changed some of her notation to one that I find easier to interpret; $\pi_i(s)$ is i's payoff (ignoring any concern with social norms) in the game, k_i a number that reflects how sensitive i is to the relevant social norm, and $\pi_{\mathrm{ml}}(s)$ is the maximum loss to any individual resulting from any norm violations by individual strategies in the strategy profile s, but never less than zero. How badly someone violates a norm is measured by the maximum impact the norm violation has on any player's preferences. If there is no loss to the norm violation or no norm violation, then $U_i(s) = \pi_i(s)$; k_i is specific to the particular situation. Agents may be intensely concerned to adhere to some norms in some situations and unconcerned to adhere to others in other situations.

To illustrate how this is supposed to work, consider the following prisoner's dilemma game that i and j would play if they were not influenced by any social norms (53–54):

	j	
i	C	D
C	2, 2	0, 4
D	4, 0	1, 1

The numbers here are interval-valued indices of preference. Suppose there is a social norm calling for both players to cooperate (play C). If both play C, then the second term in the utility function for both is zero, so the payoff for the strategy pair (C, C) is unchanged. If i plays C, and j violates the norm and plays D, then $U_i(\mathrm{C}, \mathrm{D}) = 0 - 2k_i$ and $U_j(\mathrm{C}, \mathrm{D}) = 4 - 2k_j$. Similarly, $U_i(\mathrm{D}, \mathrm{C}) = 4 - 2k_i$ and $U_j(\mathrm{D}, \mathrm{C}) = 0 - 2k_j$. Since no norms are applicable to i or j when both play D, the payoffs to the strategy pair (D, D) are unchanged. So the actual payoffs in the game i and j are playing are as follows:

	j	
i	C	D
C	2, 2	$-2k_i$, $4 - 2k_j$
D	$4 - 2k_i$, $-2k_j$	1, 1

If both k_i and k_j are greater than or equal to one, then this is a coordination game. If i and j are norm followers, then they believe the other will play C, and both will consequently prefer to play C as well.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

Several features of this construction are noteworthy. First, the extent to which individuals care about adhering to a social norm depends only on the maximum loss to any individual from anybody's norm violation, not on how many individuals will lose. Suppose that in a 10-person game, j has a choice among three strategies. The first conforms with the norm. The second causes a loss of 3 to one individual, while the third causes a loss of 2 to six players. On Bicchieri's account, the second strategy rather than the third is the more serious norm violation, and for the right value of k_j and the right payoffs, it might be rationally justifiable for j to violate the norm by playing strategy 3, but not by playing strategy 2. This oddity, which cannot arise in two-person games, seems indefensible. The problem can also be fixed easily. Just replace $\pi_{\text{ml}}(s)$, the maximum loss from norm violations in s, with the sum or average of the losses from norm violations.

Second, notice that Bicchieri's construction requires that payoffs be interval-valued and interpersonally comparable. One can sensibly talk about the maximum utility loss from any norm violation only if each individual's utility is measured on the same scale. This feature of Bicchieri's account does not fit comfortably with her account of the motivation to do what others expect, and it seems to me an unreasonable idealization to suppose that interpersonally comparable interval-valued payoffs are common knowledge.

A third implausible feature is that the effect of a loss due to a norm violation on j's utility does not depend on whether it is j who is the loser. Consider, for example, a \$10 ultimatum game between two people who are equally sensitive to a norm calling for an offer for \$5. Suppose that their utilities before modified by their sensitivity to norms are linear in the amount of money they get. As a "a fairly stable disposition" (116), an individual's sensitivity to the fair-division norm should not depend on whether he or she is the proposer or the respondent. So making an offer of \$3, for example, the proposer's utility is $7 - 2k$, while that person's utility if a respondent would be $3 - 2k$. But intuitively the norm violation matters more to the respondent than to the proposer.

A fourth feature of this account, which Bicchieri does not emphasize, is that the payoff function, $\pi_i(s)$, may be sensitive to many things besides the material payoff to a person. It could reflect altruism, benevolence, inequality aversion, personal norms, and so forth. So Bicchieri's account might complement rather than compete with other explanations of the behavior that economists observe in the laboratory. Of course, her theory will be more powerful if incorporating social norms replaces rather than adds to other factors invoked to explain behavior, but Bicchieri need not deny that factors such as altruism influence the behavior of some experimental subjects. And since it seems obvious that other factors do influence behavior, that is a good thing.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

3. Application to Ultimatum Game Experiments and Some Anomalies. In so-called ultimatum game experiments, experimental subjects, who remain anonymous to one another, are assigned the role of either proposer (P) or responder (R). P proposes a distribution of some sum of money (which in this discussion will be \$10) and R accepts or rejects the proposed distribution. If R accepts, each receives the sums specified by P. If R rejects the proposal, both get nothing. Suppose that the utility functions of P and R (apart from their concerns with social norms) are linear in money. One can then incorporate the influence of norms as Bicchieri suggests. To simplify, I shall suppose that P and R agree on the social norm, which calls on P to offer R some proportion $q \leq .5$ of the \$10. So both agree that the norm calls for an offer of \$10q. On the simplifying assumption that P never offers more than the social norm requires, the utility functions are:

$$U_{\mathrm{P}}(x) = 10 - x - k_{\mathrm{P}}(10q - x) \text{ if accepted and } -k_{\mathrm{P}}(10 - 10q) \text{ if rejected.}$$
$$U_{\mathrm{R}}(x) = x - k_{\mathrm{R}}(10q - x) \text{ if accepted and } -k_{\mathrm{R}}(10 - 10q) \text{ if rejected.}$$

Bicchieri disagrees. She maintains that if the offer is rejected, the utility for both is zero (108). This assumes that no norms apply when the offer is rejected. But in rejecting the offer, R is acting on a strategy that is in accord with the norm. So the norm applies to P.[1] If R were not acting in accord with the norm, then $U_{\mathrm{R}}(x) = x$, and R would accept any positive offer. So when R rejects P's offer, the norm applies to P, and P violates the norm, costing herself $10 - 10q$ and costing R $10q$. Since $q \leq .5$, $10 - 10q$ is the maximum loss from the norm violation.

R will then accept the offer if and only if $x - k_{\mathrm{R}}(10q - x) \geq -k_{\mathrm{R}}(10 - 10q)$, that is, if and only if $x(1 + k_{\mathrm{R}}) \geq k_{\mathrm{R}}(20q - 10)$. If $q \leq .5$, the right-hand side is never positive, and R will accept any offer. So on this interpretation of the theory, it fails to capture the phenomena.

Let us instead follow Bicchieri and suppose that both $U_{\mathrm{P}}(x)$ and $U_{\mathrm{R}}(x)$ are zero when P's proposal is rejected. In that case, R will accept the proposal if and only if $x - k_{\mathrm{R}}(10q - x) \geq 0$, that is, if and only if $x \geq 10q[k_{\mathrm{R}}/(1 + k_{\mathrm{R}})]$. The larger the k_{R}—that is, the more sensitive R is to the social norm—the closer x needs to be to the norm in order for R

1. For example, consider a simpler game, in which P has a choice of two (pure) strategies: to offer an equal division or to offer a (\$9, \$1) division. R has a choice among four strategies: (i) accept both offers, (ii) reject both, (iii) accept (\$5, \$5) and reject (\$9, \$1), or (iv) reject (\$5, \$5) and accept (\$9, \$1). The second and fourth strategies and perhaps the first strategy violate the cooperative equal-division norm, while the strategy of accepting the (\$5, \$5) division and rejecting the (\$9, \$1) division adheres to the norm. On the assumption that those who reject an offer are following a strategy such as this one rather than a strategy calling for the rejection even of equal offers, responders are adhering to the norm.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

to accept the offer. Even if P is completely insensitive to the social norm, P may make an offer that is close to the norm if P suspects that R is sensitive to the norm. If one applies the model this way, it can apparently explain why a wide variety of offers are made, though virtually never offers of more than \$5, and why low offers are rarely made and often rejected. It explains why lower offers are more readily accepted when the offers are generated by a random process than by an agent in the role of a proposer, and it explains many other specific findings.

But I've got my qualms. First, as we have seen, the application is questionable. Second, unless one maintains that q, the fraction that the norm specifies should be offered to R, depends on the quantity of money to be divided, Bicchieri's model implies that behavior in ultimatum games should be independent of whether the amount to be distributed is \$10 or \$10 million. In fact, there is evidence that play in ultimatum games is much the same whether the amount is \$10 or the equivalent of perhaps \$5,000. But how many people in an ultimatum game played for \$10 million will turn down \$500,000 because the proposed distribution is so unequal? For obvious reasons the experiment has not been done. But it is hard to believe that unequal divisions will be turned down when the sums are extremely large.

This alleged fact does not refute Bicchieri's model, because she can maintain that different norms apply when the sums are large or that people's sensitivities to social norms fade as the sums grow large. Both responses seem ad hoc. Some other way of making preferences depend on commitments to social norms is needed that allows more explicitly for ways in which opportunistic motivations grow relatively stronger as personal benefits increase.

Other experimental results also cast doubt on Bicchieri's theory. Güth and Van Damme (1998) investigated a variation on the ultimatum game involving an additional player. P suggests an allocation of some sum such as \$10 to all three of the players—P, R, and D—and R accepts or rejects the proposal. If R accepts, the money is distributed as proposed. If R rejects the proposal, the players get nothing. The third player (D) is a dummy, who makes no choices. In Güth and Van Damme's experiments, neither P nor R shows much concern about D's share, which is low and declines as players gain experience. In a variant in which R was informed only about how much P proposed to distribute to R, the amount allocated to the passive third player, D, was on average about 5% of the total. Proposers seem to feel no obligation to allocate much of anything to D. Nor do responders feel any obligation to protect D's interests. Indeed, Güth and Van Damme maintain that no rejection could be definitely attributed to a concern about D's share. The share allocated to D is considerably less than the amount given in so-called dictator games.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

If there is a social norm of fair division governing this three-person interaction, then one would expect P to propose to R that D be offered a substantial share, and one would expect R to reject offers when the proposed allocation to D was too low. How then can one square the experimental results with Bicchieri's theory? To explain what actually happens, Bicchieri would have to say either (i) that no norm governs the three-person interaction, (ii) that the norm that governs does not require any consideration of D's interests, or (iii) that something about the three-person interaction or the social norm that governs it makes people particularly insensitive to the norm. It seems to me that the most plausible response is the second one: in the three-person variant there is a norm requiring a generous share to R, with no requirements concerning D's share.

This response, like the other two, seems ad hoc. It seems that one is merely transcribing a description of how experimental subjects behave into one's list of social norms. Furthermore, some features of the data seem to rule out this explanation. In one variant of the game, it was common knowledge that R would be informed only of the amount of money that P proposed to distribute to D. If the social norm governing the three-person game did not require anything to be distributed to D, one would expect nothing to be offered. In fact, about 20% of the total was offered to D in this information condition. Furthermore, it seems doubtful that there is any effective norm calling for sharing between P and R, because proposed allocations to responders declined from roughly a third in the circumstances where R was informed of his or her own allocation to 10% or less in cases where R was informed only of the proposed allocation to D.

What these data suggest is that there is a social norm of equal division, but it seems to motivate strongly only those who are getting the short end of the stick. Furthermore, although there is evidence that people do interpret experimental circumstances in the light of their previous experience and attempt to follow established scripts, it seems that subjects are also able to appreciate strategic nuances and abandon social norms when they can get away with it. Since the subjects in Güth and Van Damme's experiment were not made anonymous to the experimenters, their concerns about the possible normative expectations of the experimenters were not very forceful. It seems either that some other model is needed concerning how social norms influence preferences or that social norms are unimportant factors in experiments such as these.

4. Fairness and Social Norms. To say that there is a social norm N prescribing certain behavior in a situation S is not by itself to say anything about whether the prescribed behavior is fair or not. There is, for example,

https://doi.org/10.1086/594529 Published online by Cambridge University Press

a social norm requiring that one wear dark clothes to a funeral, but whether one conforms has nothing to do with fairness. Indeed, there are racist and sexist norms prescribing unfair treatment of minorities and women. Yet people may regard what is customary as fair, and many of the norms that Bicchieri is concerned with—particularly those governing behavior in the ultimatum game—seem intuitively to be norms of fairness. Bicchieri speaks repeatedly of fairness norms, and one of her objectives is to understand how people form fairness judgments (100). She contrasts the fairness claims of equal shares or of unequal but equitable shares that are justified with reasons such as desert or need. Yet she never presents an account of what fairness is or what distinguishes a fairness norm from a social norm that is not a fairness norm. Since her objective is entirely to contribute to positive theory and not at all to normative theory, perhaps this does not matter. But if it is the case that people's recognition of some behavior and norms as fair and others as unfair matters, then something has been left out of her positive theory.

Although Güth and Van Damme's results seem at first glance to suggest that concerns about fairness have little influence on people's behavior, in fact, they suggest several ways in which moral appraisal of actions matters. In the circumstances in which R is informed only of the amount of money offered to D, why should P propose a relatively generous allocation for D? Presumably proposers offer a good deal more to D when they know that R will be informed only of how much D is offered because they believe that generous offers to D will increase the probability that responders will accept the proposed division. Why should they believe this? One possibility is that proposers believe that responders dislike inequality. Holding fixed his or her own unknown share, the overall distribution may be less unequal if D gets more. But inequality aversion doesn't seem to have much effect in the other information conditions, and this explanation is not very plausible. A more plausible view is that by offering more to D, proposers hope to convince responders that they are unselfish "good guys" who have been generous to R as well. But the more that is allocated to D, the less there is for P and R to divide. For R to take a generous allocation to D as a reason to accept the proposal (and for P to expect this) seems to rely on a shared moral appraisal of actions and characters.

Why are people willing to give up substantial amounts of money when they would get the short end of the stick, but not when others are treated meanly? My conjecture is that responders *resent* unequal offers to themselves and are willing to make a financial sacrifice to punish proposers for making such offers. Although certainly not an impartial reaction, what motivates rejections seems to be a moral reaction: responders turn down the unequal offers because they feel that they are not being treated as they ought to be treated. It is not just that P's actions violate a social

https://doi.org/10.1086/594529 Published online by Cambridge University Press

norm and do not meet R's expectations—norm violations and unexpected actions do not generally give rise to resentment. A highly unequal offer to R does not just violate a social norm—a highly unequal offer to D does that as well. In addition, a highly unequal offer to R fails to treat R with respect. It shows a lack of consideration for R's interests. Its unfairness to R is what leads R to reject it. Rather than a herd instinct, a desire to avoid displeasing others or the recognition of the legitimacy of the expectations of others, I see resentment and anger at work here, which depend on moral views concerning respect and fairness.

Fairness matters in ways that have not been captured in Bicchieri's theory of social norms. Attitudes or sensitivity toward norms, which Bicchieri attempts to capture with her constant k, depend among other things on people's moral appraisal of the behavior called for by norms and of norm violations. Consider another feature of Bicchieri's theory. R's utility from accepting an offer of x (in a division of \$10 with a norm of \$$N$) is $x - k_R(N - x)$. Since many offers of \$5 are made, it seems that many people believe that an equal-division norm applies and that many people are relatively sensitive to the norm. Since in fact offers of \$4 are typically accepted, it must be the case either that many people think the norm is to offer \$4 rather than \$5 or that k_R is usually less than 5.

There is nothing unreasonable about these possible explanations within Bicchieri's framework. But it seems to me that there is a more plausible alternative. An equal division is fair whether or not it is a social norm in Bicchieri's sense. Minor deviations are tolerated by respondents, since they do not expect moral perfection of one another and rejection is costly. A major deviation, however, fails to show respect. It fails to treat the respondent as an equal. And it provokes a rejection. The fact that turning down a \$4 offer is more costly than turning down \$1 has, I suspect, less to do with the behavior than the insult that is implicit in a \$1 offer. I conjecture that in many instances the relationship between the inequality of offers and the rate of rejection is explained by where people draw the line between, on the one hand, some chiseling and, on the other, a lack of respect, rather than by sensitivity to a social norm.

5. Conclusion. Both Bicchieri's theory of social norms and her view of how social norms can influence preferences are major contributions. I've expressed serious doubts about the latter, but it is a serious effort to tackle a major theoretical problem. Although there is, I maintain, more to fairness than can be captured by Bicchieri's account of social norms, her account is an important achievement.

https://doi.org/10.1086/594529 Published online by Cambridge University Press

REFERENCES

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

Güth, Werner, and Eric van Damme (1998), "Information, Strategic Behavior, and Fairness in Ultimatum Bargaining: An Experimental Study", *Journal of Mathematical Psychology* 42: 227–247.

https://doi.org/10.1086/594529 Published online by Cambridge University Press