ELIZABETH ANDERSON

Beyond Homo Economicus: New Developments in Theories of Social Norms

For more than a century, *Homo economicus* has exclusively populated the theoretical world of economics. This model of the rationally self-interested actor has also come to dominate substantial subfields of political science, sociology, law, and philosophy. However, many theorists doubt whether this model can explain most social phenomena unless it is supplemented with more socially sophisticated elements, such as social and ethical values, altruism, and desires for social status. Among these theorists are Avner Ben-Ner and Louis Putterman, who have published the results of such supplementation by various contributors in *Economics, Values, and Organization.* The contributors ask: Why and when do people cooperate? How do social norms evolve? How do values and incentives interact and influence social organizations and market outcomes?

These questions lead to one of the central puzzles of social theory: that of explaining why people comply with social norms.[1] A social norm is a standard of behavior shared by a social group, commonly understood by its members as authoritative or obligatory for them. Social norms differ from moral norms: they need not have moral content or be viewed as morally obligatory (consider norms of fashion). Nor are they the same as norms of rationality, which apply to the individual as such,

A review of Avner Ben-Ner and Louis Putterman, eds., *Economics, Values, and Organization* (Cambridge: Cambridge University Press, 1998) (henceforth referred to as *EVO.* I thank Don Herzog, Jane Mansbridge, and the Editors of *Philosophy & Public Affairs* for helpful comments on this paper.

1. This essay largely sets aside related grand problems of social theory, such as explaining the content of social norms or why they change over time.

 Philosophy & Public Affairs 29, no. 2

regardless of group membership. Yet, they still appear to be backed by some kind of normative force. Let us call this understanding of group members that they all *ought* to obey the standard of conduct defined by a social norm the *normativity* of the norm. Social norms are also typically enforced by sanctions such as praise and blame, social inclusion and exclusion. The normativity of the norm is whatever members of the social group appeal to in holding one another accountable to it and justifying the imposition of sanctions.

Social theory today offers three broad strategies for explaining why people comply with social norms. (1) Rational choice theory uses the model of *Homo economicus*. It explains behavior in conformity with social norms as the product of the strategic interactions of instrumentally rational, self-interested individuals. (2) Evolutionary theory uses models of biological or cultural evolution. It explains conformity to social norms as the expression of heritable genetic or cultural traits that have differential success in replicating themselves due to some selective process. (3) The third explanatory strategy uses models of *Homo sociologicus*—what I shall call here "social" or "cultural" rationality. It explains conformity to social norms in terms of the normativity of norms, and grounds that normativity in the ways individuals see norms as meaningfully expressing their social identities, their relationships to other people, or shared intentions and values.[2]

These three explanatory strategies each bear a different relation to the point of view of the agent. Evolutionary theory takes a point of view external to the agent. A norm could spread because of selective pressures that work independently of whatever their adherents see as binding them to obey it. Social rationality explains social norms from the adherents' own point of view. On this view, most people have *internalized* the norm and will obey it because of its normativity, apart from the sanctions attached to it. Rational choice theory represents individuals as taking a more alienated posture toward social norms. Although they may see that general conformity to a norm would be desirable, this does not provide them with a reason to conform, so long as personal conformity is, on net, costly to each agent. Only incentives contingently attached to the norm could provide a rational, self-interested individual a reason to conform. A person's reasons for conformity are thus *external*

2. I shall explain below how these ideas are related in one account of social rationality.

to the normativity of the norm, incidental to whatever might make its adherents approve of general conformity to it.

Amartya Sen argues that these three explanatory strategies are complementary, not mutually exclusive.[3] A norm could be followed both because of its perceived intrinsic merits and because of incentives. Its perceived intrinsic merits could include both prudential and impersonal goods. It could even have been established by the forces of both mindless selective pressures and deliberate institution. Sen is correct. But his bid for peace among competing schools of social theory obscures dependency relations among the different explanatory strategies. I shall argue that the normativity of norms plays an indispensable role in accounting for the motive to comply with them. Rational choice explanations of norms are dependent on social rationality. This conclusion draws upon the theories and evidence provided by *Economics, Values, and Organization*, while pressing most of its contributors to be even bolder in challenging the model of *Homo economicus*.

I. Empirical Evidence about Human Motivation

Orthodox rational choice theory attempts to explain social outcomes by assuming only the characteristics of *Homo economicus*: instrumental rationality and self-interest. Cast as methodological principles, these assumptions have considerable appeal. Methodological rationalism—the principle that we should try to explain people's actions as rational before resorting to explanations that represent them as irrational—is a sound starting point for social theory. The widespread normative appeal of the economic theory of rational choice therefore supports its use as the default theory for explaining human behavior. The theory can also be axiomatized and facilitates formal, quantitative modeling of human behavior. Methodological egoism—the principle that we should try to explain people's actions as self-interested before accepting their typically more flattering self-representations—supports the critical, unmasking function of social theory. Also, given that self-interest is one of our primary motives, a theory that could explain all human behavior without resort to other motivations could lay a claim to greater parsi-

3. Amartya Sen, "Foreword," in *EVO*, pp. vii–xiii.

mony. How far do these assumptions advance our understanding before we must resort to alternative explanations?

With respect to the hypothesis of expected utility maximization, the answer is: not far. We are not very good at judging probabilities; we do not think about risks in the way decision theorists think we ought; we do not order our preferences consistently; we care about sunk costs; and we systematically violate just about every logical implication of decision theory.[4] There is probably no other hypothesis about human behavior so thoroughly discredited on empirical grounds that still operates as a standard working assumption in any discipline. This is not for lack of alternatives. Theories of bounded rationality, prospect theory, social rationality, and other alternatives are on hand.[5]

The contributors to *EVO* are more concerned with the self-interest hypothesis. A person's motive is self-interested only if she seeks a goal *out of love for herself. Any* action-causing attitude toward *any* person or thing other than love for oneself—even if it happens to advance one's self-interest—is a motive distinct from self-interest. What other motives do we have besides self-interest? Jane Mansbridge argues for three basic human motives: self-interest, love of others, and duty.[6] Ben-Ner and Putterman advocate a similar scheme of self-regarding, other-regarding, and process-regarding preferences, except that they recognize that others may be regarded unfavorably.[7] Their scheme recognizes that actions out of hatred toward others are not self-interested. People may, out of hatred, accept their own destruction in the process of destroying

4. Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982); Daniel Kahneman and Amos Tversky, "Choices, Values, and Frames," *American Psychologist* 39 (1984): 341–50; Daniel Kahneman, Jack Knetsch, and Richard Thaler, "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy* 98 (1990): 1325–48; Richard Thaler, *The Winner's Curse: Paradoxes and Anomalies of Economic Life* (New York: Free Press, 1992).

5. Mary Zey, ed., *Decision Making: Alternatives to Rational Choice Models* (Newbury Park: Sage Publications, 1992); Herbert Simon, *Models of Bounded Rationality* (Cambridge, MA: MIT Press, 1982); Daniel Kahneman and Amos Tversky, "Prospect Theory: Analysis of Decision Under Risk," *Econometrica* 47 (1979): 263–91; Martin Hollis, *The Cunning of Reason* (Cambridge: Cambridge University Press, 1987).

6. Jane Mansbridge, "Starting with Nothing: On the Impossibility of Grounding Norms Solely in Self-Interest," in *EVO*, pp. 151–68.

7. Avner Ben-Ner and Louis Putterman, "Values and Institutions in Economic Analysis," in *EVO*, pp. 20–22.

their enemies. All three authors rightly distinguish fair dealing from love of others. One can, out of a sense of duty, pay back one's debts to someone one hates.

A more complete scheme of motives would recognize several additional facts. First, almost any attitude toward others can also be self-directed, even if it is negative. One may undermine one's self-interest out of self-loathing or shame. Second, we have other favorable self-regarding attitudes that sometimes conflict with self-love. Out of pride, some people refuse charity, even though this brings them to ruin. Third, almost any attitude toward self or others can be taken toward groups including oneself. Thus arise various "we-regarding" motivations, such as ethnic pride and shame, patriotism, and friendship. These motives do not sharply distinguish the welfare of the agent from the welfare of the group.

Still other motives are not oriented toward persons or their welfare at all. People act out of regard for animals, plants, and things. Some have high achievement motivation—the desire to excel in practices such as athletics, music, and the professions—which they pursue at the expense of self-interest. We share with animals various appetites and drives such as hunger, thirst, and curiosity. These motives need not advance the individual's interests. Curiosity killed the cat, and our health suffers today from what may be a genetically programmed desire for fatty foods. Whatever their effects, such motives cannot be self-regarding because they are also possessed by animals too primitive to have a sense of self.

This incomplete catalogue of motives offers numerous conceptual alternatives to self-interest. Social scientists have developed sophisticated empirical tests capable of distinguishing self-interest from such motives, to the detriment of the self-interest hypothesis. Mansbridge discusses an ingenious series of experiments, designed by the psychologist Daniel Batson, that elicited helping behavior from strangers in contexts that ruled out nonaltruistic explanations for it—for example, the desire to avoid feelings of guilt, to avoid distress at seeing others suffer, or to obtain social approval.[8]

Social scientists have focused more intensively on cooperation, fairness, reciprocity, and trust than altruism. Ernst Fehr and Simon Gächter

8. Mansbridge, "Starting with Nothing," pp. 158–59.

discuss experiments, using real monetary payoffs, in which subjects engaged in reciprocal cooperation (tit-for-tat) in contexts in which cooperation and retaliation came at a net cost to the agent, and without opportunities to build a reputation for reciprocity.[9] Their results contradict the rational choice prediction that people will never live up to, and, knowing this, will never offer, incompletely enforceable contracts. Fehr and Gächter found that people playing the role of firms commonly offer wages considerably higher than the returns they could expect from the enforceable level of worker effort. People playing the role of workers respond to this generosity by making considerably higher efforts than the minimum required to obtain their promised wage. Aggregate effort levels are lower than what firms demand, indicating that self-interest plays some role in determining contractual compliance. But they are much higher than what self-interest alone would produce. Firms punish shirking and reward diligence when given the opportunity to do so, even when this is costly and there is no prospect of further interaction with their workers.

These results show that people are willing to cooperate, reward cooperation, and punish uncooperative behavior, even when it is not in their self-interest to do so, and that they correctly believe that others are willing to do the same. Fehr and Gächter explain their results by postulating that people respond not just to bare incentives but to their interpretation of the intentions others express toward them in offering incentives. Generous and fair-minded intentions elicit generous and fair-minded behavior.

Trust appears to be a key factor behind the willingness to cooperate. The norm of trust tells people to act as if they believe others will reciprocate their own cooperation. It is expressed in a persistent willingness to put oneself at risk, even in the face of short-term losses due to failures to reach cooperative equilibria with one's group.[10] Under what conditions are people willing to act on trust—to put forward costly efforts in the hope that others will too, when high efforts from the group are the only way to achieve high gains for the group? Andrew Schotter's experimental work shows that this willingness depends on the group's prior history of cooperation or failure to cooperate, on the ease of coordinat-

9. Ernst Fehr and Simon Gächter, "How Effective Are Trust- and Reciprocity-Based Incentives?" in *EVO*, pp. 337–63.

10. Jonathan Baron, "Trust: Beliefs and Morality," in *EVO*, pp. 408–18.

ing everyone's actions around a high-effort equilibrium in multiple-equilibrium games, and on the vulnerability of individuals to steep losses if others fail to put forth comparable efforts.[11]

Fehr, Gächter, and Schotter's work contributes to an expanding volume of experimental evidence that, under a wide range of conditions, people act cooperatively, and obey and enforce norms of fairness even against their self-interest.[12] Despite this, *Homo economicus* remains a dominant framework for explaining norms. So let us consider the theoretical prospects for generating a cooperative social order on the basis of rational self-interest alone. Most theoretical work in this paradigm builds on three basic models: (1) coordination conventions, (2) repeated game theory, and (3) sanctioning. The contributors to *EVO* explore all three models, and thus offer an excellent view of their prospects. One of the themes to emerge from my examination of their prospects is that rational choice explanations of compliance with social norms do not have sufficient generality: they explain the normative compliance of actors only in a restricted range of settings, or explain the compliance of some actors only given that most people's compliance must be explained by other factors.

II. Rational Choice Theory and Social Norms (I): Convention

The theory of conventions explains the emergence of norms in coordination games. In such games there are at least two rules such that each agent prefers that if all but one person follows a given rule, then the remaining person follow it too. The problem is to fix the rule by which everyone will cooperate. In the classic case of a convention, it must be determined whether everyone shall drive on the right or the left side of the road. Once this determination is made, either by explicit agreement or spontaneous convergence on a salient rule, the convention is very stable, because no one has an interest in deviating from it all by herself, everyone has an interest in there being some convention, and it is costly to try to change the convention once it is established.[13]

11. Andrew Schotter, "Worker Trust, System Vulnerability, and the Performance of Work Groups," in *EVO*, pp. 364–407.

12. See, for example, Robyn Dawes, Alphonse van de Kragt, and John Orbell, "Cooperation for the Benefit of Us--Not Me, or My Conscience," in Jane Mansbridge, ed., *Beyond Self-Interest* (Chicago: University of Chicago Press, 1990).

13. David Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969).

The rational choice theory of conventions is widely regarded as a success.[14] However, most theorists regard its domain to be narrow, because the payoff structure of most social problems gives some people an incentive to deviate from the norm. Against this, Russell Hardin argues that coordination conventions offer the key to the great puzzle of how organizations get their individual members to follow organizational goals, as opposed to their own personal goals. Hardin claims that the operating rules of organizations constitute conventions, general conformity to which makes it costly for any single individual to deviate from them, and costly for any subgroup of them to change. For example, congressional committees constitute a coordination equilibrium. It is too costly for members of Congress to try to replace the committee structure with an alternative, even if the committees in place are blocking legislation that a majority wants. The coordination equilibrium is stable neither because of sanctions or external incentives, nor because members of Congress, out of public spiritedness, dedicate themselves to the mission of Congress, but simply because no one has an interest in overturning it.[15]

It is difficult to see why Hardin thinks this explains how self-interested individuals can be induced to serve organizational goals. To be sure, *given* the aim of passing laws, members of Congress will usually find that the best means of achieving this aim will be to go through the established committee structure. But this aim is given to them, not by their self-interest, but by their legislative roles.[16] If they were purely self-interested, why would they bother passing laws at all? Why wouldn't they just collect their salaries and perks? It merely defers the puzzle to reply that

14. See, however, Margaret Gilbert, *On Social Facts* (Princeton: Princeton University Press, 1989), pp. 329–67 for a searching critique of Lewis's theory.

15. Russell Hardin, "Institutional Commitment: Values or Incentives?" in *EVO*, pp. 422–23.

16. A state of affairs is in a person's interest only if someone who loved that person would want that state for that person's sake. Stephen Darwall, "Self-Interest and Self-Concern," *Social Philosophy and Policy* 14 (1997): 158–78. Most of the goals given to workers and officeholders by their roles are not like that. It would be absurd to suppose, for example, that someone who loved a congressperson would desire that many of the laws she votes for (which advance only the interests of certain subgroups) exist for *her* sake, apart from any rewards attached to *her* passing these laws. One might say: isn't the state of *her successfully passing laws* good for her? I reply: *success* in a role--achievement of its intrinsic goals--cannot be seen as advancing one's self-interest unless one judges those goals independently worthwhile. For something not worth doing is not worth doing well.

if they didn't do their jobs, the voters would turn them out of office. For why would self-interested voters turn out to vote?[17]

Hardin's model depends on the existence of coordination conventions that provide sufficient self-interested incentives for other agents to perform their institutionally given roles in sanctioning deviant agents and rewarding compliant agents. But he never manages to show how the whole incentive system can get off the ground without some agents (for example, voters) engaging in sanctioning for other than self-interested reasons. Moreover, in the classic coordination model, exemplified by the right-hand-driving norm, the equilibrium norm can be costlessly established by spontaneous convergence on a salient rule. By contrast, the establishment of organizational constitutions is costly, and requires explicit agreement. A collective-action problem therefore already had to be solved to establish the coordination equilibrium. Thus, Hardin's model covertly depends on non-self-interested motivations doing a lot of the work off-stage.

Hardin's failure to address these issues suggests that the conventional wisdom on coordination conventions is correct: at best, they explain only a small range of social norms. One might argue that cooperative behavior arises in coordination games, where it is in everyone's interest to cooperate, and persists as a habit in other contexts, because people cannot be bothered to reconsider the merits of cooperation on a case-by-case basis. This argument might be plausible if cooperation were insensitive to context. Yet it varies not only with the payoffs, but with other variables such as interpersonal trust, group identification, and vulnerability to severe loss if others do not cooperate.

III. Rational Choice Theory and Social Norms (II): Repeated Game Theory

Some rational choice theorists have claimed a promising solution to collective-action problems in repeated game theory. The main result to which they appeal is that mutual conditional cooperation (tit-for-tat) is an equilibrium strategy in indefinitely repeated two-person Prisoners' Dilemmas, provided the parties have perfect information and do not

17. On the difficulties of accounting for voting behavior in rational choice terms, see Donald Green and Ian Shapiro, *Pathologies of Rational Choice* (New Haven: Yale University Press, 1994).

discount the future too steeply.[18] Samuel Bowles and Herbert Gintis claim that such repeated games occur with unusually high frequency in "communities," which they define as social institutions with high entry and exit costs and nonanonymous interactions. These features limit migration and foster parochiality—a tendency to favor in-group over outgroup interactions.[19] Parochial communities sacrifice the gains from trade that could be achieved in open markets from interaction with strangers. They offset this disadvantage by fostering prosocial traits such as cooperativeness and honesty. Communities reduce the cost and increase the value of information on who can be trusted, thereby increasing the value to individuals of having a good reputation. Low information costs also foster social segmentation—a tendency of cooperators to play with other cooperators—which internalizes the noncontractable benefits of prosocial behavior to prosocial individuals, and excludes antisocial individuals from cooperative interactions. Repeated interaction also makes retaliation against cheaters a cost-effective strategy.[20]

Bowles and Gintis combine rational choice theory and evolutionary game theory into a theory of cultural evolution. If individuals emulate strategies that have proven successful, then the strategies that will be replicated in communities will be prosocial. However, communities will not survive if there is too much migration. To prevent this, migration must be costly. Emigration might be costly either because prospective migrants calculate that cooperative opportunities are richer inside their community, or because parochiality is a cultural (intrinsic) value for them. Bowles and Gintis appear to take the first option, arguing that

18. Michael Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987), chap. 3; Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984); Lee Alan Dugatkin, *Cooperation Among Animals: An Evolutionary Perspective* (New York: Oxford University Press, 1997), pp. 19–30. Dugatkin provides a concise and sobering summary of the main results from tinkering with Axelrod's iterated PD evolutionary model. The result stated in the text reports the implication of the folk theorems of repeated game theory that has received the most attention from social theorists. The folk theorems actually imply that there are many possible equilibrium strategies (not just tit-for-tat) in indefinitely repeated PDs, and, more generally, that numerous mutually beneficial cooperative equilibria can obtain in many indefinitely repeated games for which the rational strategy in a single play of the game would be uncooperative.

19. Samuel Bowles and Herbert Gintis, "How Communities Govern: The Structural Basis of Prosocial Norms," in *EVO*, pp. 208–9.

20. Ibid.

communities will survive because individuals need them to solve collective-action problems that arise with unenforceable contracts.[21] This supposes that there is a big difference between how much people trust one another inside and outside their communities. But Fehr and Gächter show that, at least in their Swiss subject populations, trust and cooperation among strangers in one-shot prisoner's dilemmas are remarkably common. The same results should be expected in any country whose residents are used to successfully interacting with strangers in the impersonal institutions of the market and the state.

One might argue that these observations undermine Bowles and Gintis's local thesis—that parochial communities are needed to solve collective-action problems—without touching their larger insight, that repeated game theory explains how collective-action problems are solved. Impersonal markets have trust-building substitutes for personal knowledge of one's trading partner, such as brand names and franchises. However, a collective-action problem already had to be solved before the situation of brand-name and franchised companies could be represented as an iterated game. Different franchises of the same company must already be correctly regarded as part of the same collective agency. Absent this, a customer's visits to different Burger Kings on her travels would count merely as a succession of one-shot games.

Repeated game theory with rational egoist players thus explains much less cooperation than is actually observed. Moreover, it is not evident that the representative collective-action problem is best modeled as a two-person Prisoners' Dilemma. Many social norms are enacted in situations better modeled by other iterated games, such as chicken and Hawk-Dove.[22] Rational egoists often fail to generate prosocial behavior in these games. In the Hawk-Dove game, "Hawk" represents the strategy of fighting for possession of some good, and "Dove" represents the strategy of yielding to a fight. The payoffs make Hawk the best strategy for a player if the other plays Dove, and Dove the best strategy for a player if the other plays Hawk. Robert Sugden shows that even small

21. However, on p. 224 they postulate that parochial cultural values will reduce the rate of migration and thereby reinforce the ability of communities to promote prosocial values. The implications of this alternative will be discussed in Section VII.

22. Michael Taylor, "Cooperation and Rationality: Notes on the Collective Action Problem and Its Solutions," in *The Limits of Rationality*, ed. Karen Schweers Cook and Margaret Levi (Chicago: Chicago University Press, 1990), pp. 222–40.

degrees of uncertainty about others' payoffs in repeated Hawk-Dove games often lead to a Hobbesian equilibrium of a war of all against all (everyone playing Hawk-Hawk), if rational agents are motivated only by material payoffs. He concludes that, under realistic knowledge conditions, *Homo economicus* cannot reliably sustain property conventions, which are represented as equilibria in which the possessor of a resource plays Hawk and the other player plays Dove.[23]

IV. Rational Choice Theory and Social Norms (III): Sanctions

The third main rational choice strategy for explaining why people obey social norms is to appeal to the sanctions attached to them.[24] This strategy suffers from several difficulties. (1) It appears only to defer the problem, for it is costly for agents to impose sanctions.[25] How, then, is a norm of sanctioning to be sustained? To argue that the failure to sanction must itself be sanctioned leads to an infinite regress. (2) In any event, we rarely sanction failures to sanction violations of social norms.[26] (3) Indeed, we actually observe norms against sanctioning breaches of some norms. It is considered rude to point out others' rudeness, and tattling is often frowned upon, even among those who endorse the norms tattlers want enforced. (4) The probability of getting caught and punished for disobedience bears little relation to the level of

23. Robert Sugden, "Normative Expectations: The Simultaneous Evolution of Institutions and Norms," in *EVO*, pp. 73–100.

24. Leading examples of this strategy include Michael Hechter, *Principles of Group Solidarity* (Berkeley and Los Angeles: University of California Press, 1987); James Coleman, *Foundations of Social Theory* (Cambridge, MA: Harvard University Press, 1990).

25. Little is gained by supposing that some people just like to bully others into following social norms. If it is supposed that they bully people to follow the norms because and only when violations hurt their self-interest, then it is mysterious why so many norms tell people to be polite, cooperative, and fair to everyone, rather than just to bullies. Revenge also explains little: (a) many norms benefit people from being generally observed, although the injury to anyone from a single violation is too negligible and impersonal to motivate revenge. It is absurd to exact revenge against someone for littering on public property, or even for evading taxes. When the costs of deviance are highly diffused rather than personally directed, as in most norms providing for the production or protection of public goods, revenge is not a plausible motivation for sanctioning. (b) In any event, there are norms against revenge. (c) Revenge is itself a norm-governed practice inexplicable in terms of rational choice theory. See Jon Elster, "Norms of Revenge," *Ethics* 100 (1990): 862–85.

26. For the first two arguments, see Jon Elster, *The Cement of Society* (Cambridge: Cambridge University Press, 1989), pp. 132–33.

compliance with legal norms. For example, among Swiss taxpayers, the probability of detection and the penalty tax rate do not predict the extent of tax evasion.[27]

Finally, (5) the theory supposes that acceptance of the authority or endorsement of the underlying social or impersonal point of the norm plays no motivating role in sustaining it. Since people's opinions about the general impersonal merits of a norm are supposed to be motivationally inert, it follows that even if everyone thought the norm was pernicious, they would still comply with it because everyone would be rewarding or punishing everyone else for compliance with or deviance from the norm. This is absurd. Why would people enforce a norm that no one endorses?

Such a situation is possible if everyone mistakenly believes that others approve of the norm and are ready to punish others for deviance. Timur Kuran argues that this situation sustained the Communist regimes in Eastern Europe even after nearly everyone had become disillusioned with Communism. Such situations are, like the pretense that the emperor has new clothes, highly unstable. After it became common knowledge that no one believed in Communism, the norms sustaining these regimes collapsed.[28] This implies that, under common knowledge of people's opinions of a norm, recognition of its normativity is not motivationally inert. For sanctions to get off the ground, either most people, or a core of influential people, must accept the impersonal authority of the norm and make this their reason for undertaking the costs of sanctioning. And if the normativity of the norm is a good reason for an individual to undertake the costs of sanctioning, it should also be a good reason for an individual to undertake the costs of complying with it. Sanctions therefore cannot supply the fundamental reason why people obey social norms.

How could sanctioning theory escape these five problems? Rational choice theorists have been most sensitive to the regress argument. To cut the regress short, one must identify a sanction that is either costless to the sanctioner, or automatic. An automatic sanction would not require intentional action, and might thus evade the first four problems. The only credible candidate for such a sanction is the emotions of ap-

27. Bruno S. Frey, "Institutions and Morale: The Crowding-Out Effect," in *EVO*, 452–54.

28. Timur Kuran, *Private Truths, Public Lies: The Social Consequences of Preference Falsification* (Cambridge, MA: Harvard University Press, 1995).

proval and disapproval other people feel upon observing someone obey or disobey a norm.[29]

Robert Sugden, and Chaim Fershtman and Yoram Weiss develop emotional sanctioning models of why people obey social norms.[30] To fit within a rational choice framework, such models make three assumptions. (1) People approve and disapprove of others for obeying or disobeying norms. (2) These feelings are costless and/or automatic and unintended. (3) People want others' approval.

Besides providing incentives to obey social norms, moral sentiments are needed to explain the *normativity* of norms—their being regarded as impersonally authoritative rules that everyone *ought* to obey. Sugden argues that *normative expectations* account for the normativity of norms. Normative expectations exist when people resent others for frustrating both their empirical expectations of conformity to a rule and their interests. Because people feel unease at being the target of others' resentment, normative expectations give them an incentive to conform to rules.

Emotional sanctions can arguably solve the first three problems with sanctioning theories. (Let us set aside the fourth problem, of explaining why people conform to norms even when unobserved.) The hallmark of a rational choice explanation of norms is its denial that an agent's acceptance of the normativity of a norm plays a direct role in motivating her compliance with it. Sugden's model confirms this: it is not her own approval of the norm that motivates her to conform to it, but rather *other people's* normative expectations that motivate her to conform. The fifth challenge to sanctioning theories says that this is incoherent. Whatever motivates an individual to sanction others for deviance also gives her a direct reason to comply.

So, if resentment can cause other people to comply with a norm, it can also cause the people feeling it to comply. Against this, one might argue that one would never resent *oneself* for violating a norm, because resentment is directed only at what frustrates one's own interests. However, if resentment reflected only an individual's self-concern, it could hardly function as a sanction. Sanctions are needed to motivate people

29. For this argument, see Philip Pettit, "*Virtus Normativa*: Rational Choice Perspectives," *Ethics* 100 (1990): 725–55.

30. Robert Sugden, "Normative Expectations"; Chaim Fershtman and Yoram Weiss, "Why Do We Care What Others Think About Us?" in *EVO*, pp. 133–50.

who don't already care about the interests of others. If an agent does not care about another's interests, why would she care about his self-interested feelings?

Sugden recognizes this fact and follows Hume in insisting on the impartiality of the moral sentiments. He asks, why, when traveling on a crowded train without a seat, don't I take the seat of someone who leaves it to go to the toilet? He answers, because it is not just she who would resent me for taking the seat; everyone else, accepting the convention that such seats are not to be taken, would take her side and share her resentment of me. My unease at being exposed to the impartial ill will of many other people deters me.[31] Sugden's explanation assumes that people feel resentment on behalf of others, not just themselves. But, assuming he shares the same moral sentiments as others, he should by his own lights reproach himself for taking the passenger's seat. He should not need to rely on the reproach of others to motivate him, since, given the impartiality of his moral sentiments, they can just as easily be directed against himself as against any other person. Sugden may doubt whether self-reproach could motivate people to follow cooperative norms against their self-interest. But why should the reproach of others be more powerful? "What are my feelings of self-reproach to me?" is a less compelling question than "What are their feelings to me?"[32] Emotional sanctions, no less than behavioral sanctions, fall prey to the objection that whatever motivates the sanctions can directly motivate compliance with a norm. Sanctions are only a supplementary motive to the original motive for compliance, without which the norm would never have been established: acceptance of the normativity or impersonal authority of the norm.

V. Evolutionary Theories of Social Norms

An evolutionary explanation of a trait explains its frequency in a population in terms of selective pressures that favor or disfavor its replication relative to its alternatives in a particular environment. A full evolution-

31. Robert Sugden, "Normative Expectations," p. 85.

32. One might suppose that the reproach of others would be more powerful because it causes embarrassment or a loss of status. But these facts would rationally motivate people only if they already cared about others' normative judgments about how their actions ought to affect others' interests. Why would people care about such judgments if they do

ary explanation of any trait involves three mechanisms: (a) a mechanism for generating a diversity of traits; (b) a mechanism for replicating these traits; and (c) a mechanism for selecting some traits over others to be favorably replicated, on the basis of the consequences or "payoffs" of expressing them in the environment. These mechanisms need not be biological. If cultural mechanisms could be identified that play roles (a)-(c), they would ground an autonomous theory of cultural evolution.

The marriage of evolutionary theory with game theory has raised hopes for such a theory. The key contribution of game theory to evolutionary theory is the thought that the primary determinant of the payoffs of any particular social trait is the frequency distribution of itself and its alternatives in the population. In successive rounds of the game of life, people (or their descendants) are assumed to adjust (or inherit) their social traits in response to the payoffs from expressing those traits in the previous round. This allows the creation of mathematically tractable endogenous theories of social traits.

One can create a game-theoretic model of cultural evolution that generates any social outcome one likes, if one is free to choose one's starting assumptions. To avoid a "just so" story, the assumptions of an evolutionary model should meet the following constraints.[33] (1) *Physical realizability.* The postulated selection, replication, and generation mechanisms should be consistent with human capabilities and limitations. (2) *Representativeness.* The game being played should be representative of critical, enduring, frequently encountered features of the interactive environment our ancestors faced. (3) *Robustness.* The desired social outcome should be generated under a variety of starting conditions (such as initial frequency distribution of social traits) and should not hang on implausible conditions (such as perfect information and flawless execution of strategies). Since we do not know the specifics about our ancestors' environment, the outcome should hang on its generally representative features and not on rigging the details in peculiar ways. (4) *Avoidance of anachronism.* The model should not project onto the past

not already care about others' interests? The general problem is that emotions have an evaluative basis. It is hard to credit emotions with intrinsic reason-giving force without crediting the evaluations that justify those emotions with comparable reason-giving force.

33. This list is drawn from the important article by Justin D'Arms, Robert Batterman, and Krzyzstof Górny, "Game Theoretic Explanations and the Evolution of Justice," *Philosophy of Science* 65 (1998): 76–102.

human capabilities, social settings, and problems that have emerged only recently. This is a logical implication of realizability and representativeness, but worth stating separately to remind us of the temporal dimension implicit in these criteria.

Game-theoretic models of cultural evolution differ according to the unit of selection. Some models endogenize strategies or behaviors; others endogenize underlying motivational states. When an evolutionary model endogenizes behaviors, it explains the frequency distribution of these behaviors. It does not explain their *normative* status: the fact that agents judge that people *ought* to behave in that way and hold each other to account for their behaviors. Thus, it is odd for evolutionary game theorists such as Brian Skyrms to claim that, in demonstrating that the strategy of splitting windfalls evenly among finders is an evolutionarily stable equilibrium, he is explaining "the origin of our concept of justice."[34] How does it explain the fact that we consider unequal divisions of windfalls unfair, as opposed to unexpected?[35]

A more promising strategy for explaining the evolution of norms would take some underlying motivational state as the unit of selection. A suitably selected motivational state could simultaneously supply a mechanism for norm-conforming behavior and explain the normativity of the norm. This is in keeping with the conclusion defended above, that the normativity of a norm plays an indispensable role in motivating conformity with it. Sugden, Chaim Fershtman and Yoram Weiss, and Ken Binmore all take motivational states as their unit of selection.

Fershtman and Weiss present a model of the evolution of a preference to care about the opinions of others concerning how one ought to behave.[36] Suppose people costlessly confer social esteem on those who cooperate more than average in Prisoners' Dilemmas, and pass on their

34. Brian Skyrms, *The Evolution of the Social Contract* (Cambridge: Cambridge University Press, 1996), p. 21.

35. Sugden claims that his concept of normative expectations accounts for normativity by grounding resentment, without agents' presupposing any judgments of how others ought to behave. Mansbridge objects that, unless one judges that others are *obligated* to behave as one expects, that they frustrate one's expectations will lead only to irritation, not resentment. Mansbridge, "Starting with Nothing," pp. 162–63. Sugden replies that requiring the moral sentiments to have normative content violates a naturalistic constraint on social scientific explanation. Sugden, "Normative Expectations," p. 84. But all that Mansbridge requires is that people represent the objects of their resentment as having acted unjustly, not that their representations are true.

36. Fershtman and Weiss, "Why Do We Care What Others Think About Us?"

preference for esteem to their children. Fershtman and Weiss identify conditions in which a preference for esteem can spread under natural selection, even if only material payoffs increase reproductive fitness. Those who prefer social esteem triumph because only they reap the material gains of cooperation in PDs.

Fershtman and Weiss's model has heuristic value against those who think that ruthless materialists must drive people who care about others' opinions into extinction. However, it is neither realizable nor representative, because its reproductive mechanism is implicitly asexual. (They assume that one's children will inherit one's preference for esteem. What preference would the offspring of two parents with different preferences have?) Suppose we introduced sexual reproduction, and hence sexual selection, into their model. Everyone will prefer to mate with people who care about their opinions. Social esteem based on responsiveness to others' opinions must therefore have a direct impact on one's reproductive fitness. So it is not plausible to assume that reproductive fitness depends only on material payoffs.

Ken Binmore offers an evolutionary explanation of normative preferences.[37] He argues that a fairness norm akin to the Golden Rule, as modeled by reasoning behind an original position, evolved with the human species. We apply this norm by asking how we would prefer goods to be distributed, on the supposition that we have an equal chance of being any of the parties to a distributive problem. To answer this question, each person must consult her *empathetic preferences*: would I rather be Adam in condition X or Eve in condition Y? These preferences express interpersonal utility comparisons. Individuals must come to share the same empathetic preferences if the original position device is to serve its evolutionary function as a standard for joint decision making. The rational choice of distributions is then made by maximizing expected utility as defined by empathetic preferences. The resulting norm of justice is utilitarian, but with a contractualist rather than a teleological rationale. Binmore deserves credit for recognizing the autonomy of normative motivation, instead of trying to explain normative conformity by

37. Ken Binmore, "A Utilitarian Theory of Political Legitimacy," in *EVO*, pp. 101–32. Binmore's theory, like Skyrms's, focuses more on the content of norms than on the motive for compliance with them. For another discussion of both problems, including a more thorough discussion of Binmore and Skyrms, see Peter Vanderschraaf, "Game Theory, Evolution, and Justice," *Philosophy & Public Affairs* 28, no. 4 (Fall 1999): 324–58.

appealing to self-interest or the desire for social approval. Let us test his model against the criteria articulated above.

Physical realizability. Binmore postulates that, in negotiating the just distribution of resources, the parties will bargain for shares based on their empathetic preferences. "Both players will test their recommended bargaining strategy against the[ir] empathetic preferences . . . and adjust their behavior until they reach a Nash equilibrium of their bargaining game."[38] This supposes that our ancestors were able to maximize their expected (empathetic) utilities. Only modern decision theorists are able to do this consciously. Binmore must be imagining, then, that some unconscious process steered ordinary mortals to the same results. But the evidence against the hypothesis that humans generally are "as if" expected utility maximizers is overwhelming. Binmore's selection mechanism therefore lacks a plausible embodiment.[39]

Representativeness. Binmore assumes that the original position device offers a plausible representation of negotiation over fair terms of distribution among our ancestors. This requires identifying a concrete distributive problem in which this device could have evolved as a joint decision-making rule. Binmore's choice of Adam and Eve negotiating over the terms of their marriage is anachronistic. For most of human history, women have not been parties to a marriage negotiation. They have been the objects of negotiation between their male kin and the prospective husband and his male kin. Nothing hangs on the marriage illustration, however, since Binmore's analysis would have run the same course whatever distributive problem he chose. There lies the problem. In commonsense moral thinking, we find numerous conflicting standards of local justice: need, desert, property entitlement, status, contract, first come-first serve, finders keepers, equal shares, and so forth. It is plausible to think that each of these evolved to solve specific local distributive problems: perhaps need for distribution among kin, contract for trade between alien tribes, equal shares among finders for windfalls, desert for dividing gains from cooperation among participants in joint production. Explicit contractualist and utilitarian princi-

38. Binmore, "A Utilitarian Theory of Political Legitimacy," p. 115.

39. This argument does not discredit all rational choice explanations of social phenomena. In certain institutional contexts, such as competitive markets, the results predicted by rational choice theory can be generated by only a few rational actors at the margins. The trouble with Binmore's evolutionary argument is that it requires people *in general* to be endowed with the skills and preference structure of *Homo economicus*.

ples offer standards of global, not local, justice. As such, they arrived very recently, as philosophical attempts to reconcile conflicts among local standards. To avoid anachronism, Binmore assumes that the original position device first evolved as a standard of local justice and only later became a global tool to be applied to the whole social order.[40] He therefore needs to identify a representative local distributive problem for which the use of the original position device was an evolutionarily stable solution. Marriage isn't an apt choice, but what would be?

Perhaps Binmore thinks that all the local standards of justice implicitly encode contractualist-utilitarian reasoning. This makes the variety of local standards difficult to comprehend. The core assumption of contractualist-utilitarian reasoning is that equal units of welfare count equally for decision making, no matter who enjoys them. The core assumption of local standards of justice is that how, and how much, people ought to concern themselves with the interests of others depends on their relationship to them: are they equals or unequals? Stranger, tribe member, friend, or kin?

To his credit, Binmore does try capture the realities of power differences in assigning weights to different people's utilities. But he captures these differences in the thought that the powerful doubt that the lower orders suffer as much from the same deprivations as the higher orders do. This preserves an egalitarian decision principle (each person's utilities still count the same) and packs the power differences into empathetic preferences. The powerful in most inegalitarian societies have entertained few illusions about the welfare enjoyed by the lower orders under the distributive principles they thought were just. They packed the power differences into inegalitarian decision principles that distributed goods according to social status and function rather than subjective utilities. The thought "being of inferior status, her interests count for little" is not to be modeled as the thought "she would not suffer as much from the same deprivation as I would."

Avoidance of anachronism. The idea that norms of justice are based on interpersonal comparisons of subjective utilities arises fundamentally in worldviews that assume that subjective utilities—preferences that people hold *apart from any feelings of obligation to hold them*—are the ultimate basis of human decision making. Such worldviews are of

40. Binmore, "A Utilitarian Theory of Political Legitimacy," p. 129.

recent origin, because the institutional conditions for their plausibility are modern. For most people to be in a position to develop subjective preferences covering a wide range of decision problems requires a social order in which they are free to choose ways of life that are not comprehensively defined by ascriptive social identities and social roles. This requires such conditions as a substantial economic surplus distributed by free and impersonal markets, elimination of caste, differentiation of social spheres, individual mobility between spheres, expansive rights to privacy, social norms of tolerance, and the widespread adoption of a conception of the self as the proper author of its own priorities. These are very late developments in human history. In societies lacking these conditions, the vast majority of people make claims to resources on behalf of their role-given goals and ascriptive identities. They have few subjective preferences on which to base claims, and those that exist have little standing, even in their own eyes. Interpersonal comparisons are based on judgments of the social importance of their roles and identities.

The great philosopher-economist John Stuart Mill was acutely aware of the historical contingency of "subjective utilities." That is why he wrote a book urging people to acquire them, and to create the social conditions in which everyone else could, too.[41] Today's economists assume that wide-ranging subjective utilities are a given of human nature, rather than a stunning social achievement. In evolutionary theorizing, this assumption runs the risk of anachronism. It misrepresents wide-ranging subjective utilities as the general cause of social norms and institutions throughout history rather than as the recent effect of peculiarly modern institutions.

Evolutionary theory is needed to explain how our general psychological capacities to adopt and obey social norms evolved. Whether it can explain how particular social norms came about remains to be seen. The most acute need for theories of cultural evolution is to find a realizable selection mechanism for cultural traits. Natural selection works in the long run, but it cannot explain short-run cultural change—that is, most of the change observed in recorded history. The main alternative selection mechanism used in evolutionary game theory is rational choice (as

41. John Stuart Mill, *On Liberty* (London: J. W. Parker and Son, 1859).

defined by decision theory). This is not physically realizable in the general population. Models of bounded rationality may be more promising.

VI. Social Rationality and Social Norms

The great puzzle of social norms is not why people obey them, even when it is not in their self-interest to do so. It is, how do shared standards of conduct ever acquire their normativity to begin with? Once we understand this, there is no further difficulty in understanding the motive to obey them. We obey them, because we believe that we *ought* to. We accept them as authoritative principles of action. This is the guiding idea of *Homo sociologicus*—that people obey norms because they have "internalized" them. Social rationality stakes its claim on the idea that the normativity or "oughtness" of social norms is not the "ought" of prudence. Nor is it the "ought" of morality. Many social norms, such as norms of fashion, have no moral content. So what is it? We need an account of how a social norm can provide an intelligible ground for action.

Viviana Zelizer hints at this in her discussion of the distinctions among compensations, entitlements, gifts, and bribes.[42] She argues that the relationship between the parties determines how to classify a payment from one person to another. For example, many people consider it an improper bribe when a parent pays her child to do chores. Why? A payment is a bribe when the compensation to the recipient is for a performance that is either not authorized by the relationship or required by the relationship without payment.[43] As family members and not employees, children are expected to do their part in household chores without payment. Properly socialized children in fact do so, accepting this expectation as an authoritative action-guiding norm.

Let us unpack the reasoning that could make this intelligible.[44] Suppose a child regards herself as a member of her family, and regards her family as a group of people dedicated to living together and therefore to working as a body to provide the conditions for doing so. Suppose she

42. Zelizer, "How Do We Know Whether a Monetary Transaction is a Gift, an Entitlement, or Compensation?", 329–31.

43. Zelizer, "Gift, Entitlement, or Compensation?" p. 332.

44. This follows the model provided by Margaret Gilbert, *On Social Facts*, pp. 422–24.

regards completing household chores as constituting what her family sees as some of those conditions. In taking up this understanding of her practical identity, constituted by membership in a family, and this understanding of her family's goal, she thereby regards herself as committed to that shared goal, and thus to doing her part in advancing it. Her commitment constitutes her reason for doing her chores. This reasoning nowhere appeals to the prospect of monetary payment, parental approval, punishment for failure to do chores, or any other incentive. This is a paradigm of practical reasoning. To count as a reason for action, a consideration must appeal to a person's self-understanding, not her self-interest. It must fit in to her understanding of her identity.

Most people's identities are largely, although not exclusively, constituted by their membership in social groups or collective agents. Theories of collective agency have recently enjoyed a great revival.[45] Here I shall focus on the theory of Margaret Gilbert's, which I find most promising. According to Gilbert, a social group is a "plural subject": a set of people who think of themselves as "we," and understand one another to be jointly committed to some goal, belief, or principle of action.[46] In so identifying with a group, an individual accepts responsibility for doing her part in advancing the group's goal. If the group is organized, she may find herself in a specific role within the organization, and she may thereby accept as her goals those given to her by her social role.

This supplies an elegant answer to Hardin's question of how people can be motivated to advance organizational goals. The answer does not presume that incentives and sanctions are not necessary to sustain organizations. A member's commitment to advance organizational goals is conditional on enough of the others doing their part to sustain an understanding that the members really constitute a coherent group. In the standard employer-employee relationship, the employee's commitment is conditional on the employer's playing his part, which includes paying compensation for work performed. Compensatory in-

45. Exemplary works include A. C. Baier, *The Commons of the Mind* (Chicago and La Salle: Open Court, 1996); J. Searle, "Collective Intentions and Actions," in *Intentions in Communication*, ed. P. Cohen, J. Morgan, and M. Pollack (Cambridge, MA: MIT Press, 1990), pp. 401–15; Michael Bratman, "Shared Intention," *Ethics* 104 (1993): 97–113; R. Tuomela, *A Theory of Social Action* (Dordrecht: Reidel, 1984); David Velleman, "How to Share an Intention," *Philosophy and Phenomenological Research* 57 (1997): 29–50; Margaret Gilbert, *On Social Facts* (Princeton: Princeton University Press, 1989).

46. Gilbert, *On Social Facts*, pp. 204–5.

centives may be needed to recruit willing people into organizational roles.

This account provides a more general explanation of the need for negative sanctions than self-interest theory does. They are needed to motivate those who do not fully identify with their role in a group. People may shirk not only out of self-interest, but out of identification with conflicting social roles (a man might be poorly motivated to perform tasks he regards as "women's work"), out of an inability to see oneself as absorbed in the task (a worker might find it boring), out of pride (a worker might find a task demeaning), or out of any number of other self-understandings that conflict with regarding oneself as committed to the task. Sanctions mobilize self-interest against the full range of other motivations. They also provide assurance to conditional cooperators that others will cooperate to the degree necessary to call forth their own cooperation.

This account of group identification explains the motive to comply with a norm in terms of its normativity. Gilbert defines a social convention or norm as a principle of action jointly accepted by a group as a simple fiat.[47] (A fiat is a principle regarded as authoritative in virtue of its joint acceptance, over and above whatever other justification might be offered for it.) In jointly accepting the principle of action, each member of the group regards herself as committed to doing her part in upholding the principle *with the others*. To regard *us* as being jointly committed to a principle is to regard each of us as thereby having a reason to comply, *and* to accept that everyone is accountable to everyone else with respect to compliance.[48] The normativity or "oughtness" of social norms, then, is an "ought" constitutive of commitments of collective agency. It is grounded in the perspective of collective agency, in "our" shared view of how "we" ought to behave. It is based on the fact that members accept the authority of "us" to determine how each should behave in the domain defined by the norm.

How should motivation by norms be represented in relation to other motives? Ben-Ner and Putterman, as well as Frey, represent individual values, including commitments to obey norms, as arguments in the in-

47. Gilbert, *On Social Facts*, p. 373.

48. This account does not distinguish joint acceptance from joint commitment. For an argument that commitments provide reasons for action, see my "Reasons, Attitudes, and Values: Replies to Sturgeon and Piper," *Ethics* 106 (1996): 538–54.

dividual's utility function, as just one preference among others to be satisfied.[49] This enables theorists to represent cases in which the individual chooses to sacrifice her values in favor of her interests. Timur Kuran objects that this misrepresents the distinctive motivational structure and function of values.[50] Evolution endowed us with discrete motivational systems: preferences are subject to trade-offs, whereas values (as Kuran defines them) present themselves as *obligatory* to the agent—as not subject to trade-offs. Preferences and values define *distinct* and potentially conflicting orderings within the self. Conflicts within the self are endemic, because there is no general function of the brain that reconciles all of our motivations.

Kuran's ideas point toward a more fruitful way of representing the relation of values and preferences. But he throws away his insights when he suggests that, in cases of "moral overload," when an individual's values demand the impossible, individuals deal with the prospect of guilt by just trading off the intrinsic utility of their options (as represented by their preferences) with their moral utility (as represented by their values).[51] This denies the distinctiveness of value-based motivation that Kuran was keen to uphold. It also postulates the very unified utility function that Kuran previously rejected, on the grounds that there is no brain function for reconciling our motivational conflicts. Of course there is such a brain function. It is called reason. Kuran himself invokes an implicit conception of reason in explaining intrapersonal conflict resolution. His mistake is to assume that reason must operate on the basis of a single unified preference ordering.

Gilbert's account of norms permits a more fruitful way of understanding practical conflicts and their resolution. Each of us is an individual agent, an "I" with, let us suppose, an associated partial preference ordering. Each of us is also typically a member of numerous collective agencies—many "we's"—a citizen of a state, an employee of a firm, a member of a church, a relative in a family, and so forth—each jointly committed to different goals, priorities, and principles of action, representable in part by distinct partial preference orderings. These rankings conflict, because the different collective agencies are not fully coordi-

49. Ben-Ner and Putterman, "Values and Institutions in Economic Analysis," pp. 20, 23; Frey, "Institutions and Morale," p. 440.

50. Timur Kuran, "Moral Overload and Its Alleviation," in *EVO*, pp. 231–66.

51. Kuran, "Moral Overload," p. 243.

nated with one another and do not automatically accept their members' personal priorities as inputs into joint decision making. ("This spending bill will make me rich" is not an acceptable reason a representative can give to Congress for passing a bill.) Reason resolves conflicts among these preference orderings not by weighing one priority against another, but by determining which ranking, in the given context, has authority. This view represents reason not as a scale upon which competing values are balanced, but as a judge drawing jurisdictional boundaries. Any given preference ordering prevails only within its jurisdiction—that is, only in contexts where its associated agent ("I", or this or that "we" to which I belong) has authority to decide what I should do.

Kuran recognizes that people may limit the scope of their normative commitments, thereby allowing their personal priorities to govern their choice. But he represents this as unprincipled and self-deceptive "casuistry" and "rationalization." Similarly, he views sphere differentiation—the practice of defining distinct spheres of life (public/private, work/home, market/state) in which people pursue different priorities, as just a technique for hiding our own inconsistencies from ourselves.[52] Like Binmore, he is caught in the grip of a picture of reason while blind to the social conditions for its realization. Reason does not demand that all of an individual's distinct preference rankings somehow get translated into a single ranking: only a hermit, who belongs to no groups, or a subject of a single totalitarian social group, could achieve this. If we view the function of reason not as weighing goods given to it, but as assessing the authority of action-guiding principles, a system of sphere differentiation that grants individuals discretion to pursue their personal priorities in the private sphere can be seen as rational, principled and transparent.

VII. Values, Incentives, and Markets: Normative Implications

Two tales are often told about the relation of incentive-based market institutions to values. The pessimistic tale represents capitalism as dependent upon a stock of social capital—trust, norms of cooperation and honesty, and other prosocial values—that are cultivated by nonmarket institutions such as families, neighborhoods, churches, and other com-

52. Kuran, "Moral Overload," pp. 251–55.

munities. Capitalist firms freely make use of this social stock without replenishing it, and undermine the very institutions that are responsible for society's social capital. In this tale, free markets are responsible for moral decline, anomie and loneliness, and eat away at their own foundations.[53] The optimistic tale represents capitalism as expanding the scope of cooperation and trust by enabling people to reap gains from trade worldwide, bridging parochial divisions of nationality, religion, and ethnicity. Capitalism is an engine of cosmopolitanism, cooling socially dangerous passions such as religious fanaticism, and overcoming xenophobia.[54] The impersonality, anonymity, and openness of markets to all comers is favorably contrasted with social orders in which people are tightly constrained by parochial connections and loyalties of family, ethnicity, and neighborhood.

Both stories recognize that free markets cannot function efficiently on the basis of self-interest alone. Many contracts, especially labor contracts, are incompletely enforceable. If people were not willing to work harder than self-interest required, and if employers were not willing to reward workers for such extra effort, many potential gains from trade could not be reaped. Moreover, markets are efficient only to the extent that participants accept the rules of the game. Once people extend self-interested reasoning to consider whether they should lie, cheat, and steal, market transactions become very costly or break down.

Given the dependency of markets on prosocial norms, we must ask: do markets expand the scope of these norms, or undermine them? Ben-Ner and Putterman worry that the pessimistic tale is true. Robert Frank and Bruno Frey supply some evidence in favor of this view.[55] Frank argues that many social norms, such as those against conspicuous consumption, represent collectively rational "positional arms control agreements," limiting competition for positional goods.[56] Markets can

53. Exemplary works in this tradition include Joseph Schumpeter, *Capitalism, Socialism, and Democracy* (New York and London: Harper & Brothers, 1942); Robert Putnam, "Bowling Alone: America's Declining Social Capital," *Journal of Democracy* 6 (1995): 65–78; Karl Polanyi, *The Great Transformation* (Boston: Beacon Press, 1944); Fred Hirsch, *Social Limits to Growth* (Cambridge: Harvard University Press, 1976).

54. Albert Hirschman, *The Passions and the Interests* (Princeton: Princeton University Press, 1977).

55. Robert Lane also takes a pessimistic view of markets, but more for making people subjectively unhappy (in substituting income for more important goods) than for undermining social norms. "The Joyless Market Economy," in *EVO*, pp. 461–88.

56. Frank, "Social Norms as Positional Arms Control Agreements," in *EVO*, pp. 275–95.

undermine such norms if they reward the people at the pinnacle of a positional hierarchy vastly more than those in the next tier. Frank argues that in such cases the state can supply substitutes for these social norms, such as a consumption tax.

Frey argues that providing people with material incentives to do socially desirable things sometimes crowds out rather than supplements their intrinsic motivation to do them.[57] Residents of one Swiss town identified as a potential site for a nuclear waste facility expressed *less* willingness to accept the facility in response to offers of compensation.[58] Frey suggests that this "crowding out" effect arises because the offer of incentives reduces people's sense of control over their choices and damages their self-esteem.

I suggest rather that the offer of compensation changed the perceived relationship of the Swiss government to the town residents and thus changed the practical identity they assumed in contemplating the waste facility. In asking the residents to accept the facility without compensation, the Swiss state addressed the residents as *citizens*. It implicitly asked them to frame their practical dilemma as: "what principle for siting the facility should *we* accept, given that we (Swiss citizens, considered collectively) must process the waste somewhere?" This way of framing the question precludes a not-in-my-backyard response, because it recognizes that the facility must land in someone's backyard. In offering compensation to the townspeople, the Swiss state represented their interest in a waste-free town as an *entitlement*, like a property right, and asked them their price for giving it up. It thereby implicitly asked each of them to frame their practical dilemma as: "how much is it worth to *me* (or we townspeople) to keep my town waste-free?" From that point of view it was harder to represent the siting of the waste facility in their town as desirable, because they no longer saw themselves as responsible for solving the collective problem they faced as national citizens, of finding some site for the facility. Frey's cautionary tale about market incentives teaches us that the offer of incentives may change the relations of the parties, and thereby invoke the norms and distinct preference rankings of the new relationship.

On the optimistic side, Bowles and Gintis argue that markets will not undermine the communities that nurture prosocial values, because

57. Frey, "Institutions and Morale," pp. 437–60.
58. Ibid., pp. 448–51.

communities offer opportunities for trustworthy exchanges that individuals cannot get elsewhere.[59] Fehr and Gächter's evidence, as well as our common experience in national markets, contradicts this. However, it may be true in communities that cultivate parochiality as a cultural norm, directing group members to extend their trust only to fellow members. In discouraging trust of outsiders, the norm of parochiality reduces the chances of successful interaction with them. Unsuccessful interactions with outsiders then reinforce parochiality. The norm of parochiality constitutes a vicious circle, and a collective-action problem for members from different communities.[60]

This puts a darker cast on community than Bowles and Gintis do. To the extent that community members reserve action on their prosocial norms to fellow community members, they will not be able to function adequately in their roles in impersonal institutions of the market and state. Susan Rose-Ackerman argues that this explains why, in many developing countries, bureaucratic corruption is pervasive, and entry into particular markets is limited to those with personal connections. The obligations of personal relationships trump those of the principal-agent relationship. Nepotism is the norm, and bureaucrats readily accept bribes to ignore the rules, or even insist on bribes just to do their jobs.[61] A solution to such problems cannot be found simply by appeal to self-interest and instrumental rationality. As Rose-Ackerman and Viviana Zelizer stress, what needs to be reformed in these cases is people's understanding of the meaning of their relationship to those with whom they are interacting. If corruption is to be reduced, bureaucrats must come to regard themselves as agents of the state and as public servants, rather than as private owners of public services for sale to the highest bidder, or as standing in a client-patron relationship to people who demand their services.

Rose-Ackerman does not advocate a straight imposition of Western norms of impartiality on developing countries. Such norms hold out the prospect of greater gains from trade, but they conflict with forms of social relationship that their members hold dear. Because these forms of social relationship are not valued merely instrumentally, the instru-

59. Bowles and Gintis, "How Communities Govern," pp. 206–30.
60. Baron, "Trust: Beliefs and Morality," p. 417.
61. Susan Rose-Ackerman, "Bribes and Gifts," in *EVO,* pp. 316–24.

mental superiority of impartial norms is not likely to sway people. If communities are resilient, it is not because they successfully compete with markets, as Bowles and Gintis suppose, but because they offer goods, constituted by nonmarket social relationships, distinct from those supplied by markets.

Thomas Weisskopf and Nancy Folbre offer a similarly measured evaluation of markets, avoiding both the optimistic and pessimistic tales.[62] They assess the consequences of substituting market provision of dependent care services for care services provided by women to their families. Pessimists, casting this trend as a substitution of self-interested motivation for direct caring, see a fatal dissolution of community and family bonds. Market-oriented optimists join feminists in celebrating this trend as a liberation of women from patriarchal coercion and as expanding their opportunities to participate in the public sphere. Weisskopf and Folbre criticize both the coercive context of patriarchal caregiving and the market's dependence on self-interest. Caregiving provided out of direct concern for those cared for is more intrinsically desirable and more reliable than caregiving out of either pure self-interest or coercion. Neither incentives nor coercion are able to increase the supply of care services *from caring motives.* Nor is a more vigorous inculcation of feminine norms of caring desirable, for it only leaves women more vulnerable to exploitation. The most just way to increase the supply of caring labor would be to take the gender out of norms of caregiving. Men are reluctant to provide much caring labor, because this is labeled "feminine." If norms for caregiving were degendered, men would be more willing to do their share.

Weisskopf and Folbre offer an attractive model of how to think about markets and values, both from a positive and a normative point of view. They integrate game theoretic insights (in representing the conflict between men and women over who should provide caring labor as a kind of chicken game) with a recognition of the autonomy of social norms (in representing their content and supporting motivation as not mere creatures of self-interest, but bound up with culture and group identities). They also recognize the genuine advantages of markets with their limitations.

62. Folbre and Weisskopf, "Did Father Know Best? Families, Markets, and the Supply of Caring Labor," in *EVO*, pp. 171–205.

VIII. Conclusion

Economics, Values, and Organization signals an exciting breakthrough for the branches of social theory influenced by economics. It demonstrates how being open to a richer representation of human motivations enables more adequate accounts of market phenomena, and advances our understanding of the relations of markets to other social phenomena.

We can also learn an important methodological lesson from *Economics, Values, and Organization.* Throughout this review, especially in my discussion of repeated and evolutionary game theory, I have pointed out places where the unrealistic assumptions of various models lead to untenable explanations of social phenomena. These problems can be traced to the methodological advice of Milton Friedman, who counseled economists to ignore empirical investigation of the actual causes of human behavior, and content themselves with theorizing on the assumption that people behave "as if" they were self-interested utility maximizers.[63]

Imagine if Friedman were a biologist. "Don't bother trying to figure out how DNA actually works," he would say, "just stick to the idea that genes act 'as if' they are selfish, and you will be able to explain just as much." If biologists had taken this advice, they would have little more to show for themselves than a collection of just-so stories about how life evolved. Friedman's advice confuses heuristics with science. To have a chance at identifying the actual causes of social phenomena, theorists need to avoid the temptation to construct just-so stories and try to square their assumptions with empirical evidence about human motivation, capacities, and circumstances. As more such evidence emerges in psychology, sociology, and history, economically inspired models can only profit from attending to it.

Some of the contributors to *Economics, Values, and Organization* have taken major steps in overcoming Friedman's bad advice. Let us hope that economically inspired social theorists follow their example.

63. Milton Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics* (Chicago: University of Chicago Press, 1953).