Virtus Normativa: Rational Choice Perspectives†

Philip Pettit

Norms are an important species of social institution, on a par with conventions, customs, laws, and other brands of established regularity. They often overlap with those other institutions, so that the same regularity can be both a norm and a law, for example. But still, they retain a distinctive profile. Like the other institutions norms reinforce certain patterns of behavior, but they do so in their own way, by representing those patterns as peculiarly desirable or obligatory. Norms are generally operative, for example, in supporting patterns of behavior like truth-telling, promise-keeping, and abstinence from theft, fraud, and violence. They also play a role in supporting familiar virtues like loyalty, fairness, integrity, and courtesy, as indeed they play a role in supporting less attractive dispositions like conformism and vengefulness.[1]

Most visions of the good society allot an important job to certain norms, relying on their presence to generate or reinforce crucial features of the society: features such as the dedication of public officials to the common interest, the acceptance by people at large of certain sorts of official decision, their participation in different forms of social and political

* The articles for this symposium were generated by a conference on norms, held at the University of Chicago, May 19–21, 1989, with support from the National Endowment for the Humanities, the Center for Ethics, Rationality and Society at the University of Chicago, and Jerry Knoll of Washington, D.C.

† I am grateful for helpful comments received from Alan Bellett, John Braithwaite, Geoffrey Brennan, Bob Goodin, Alan Hamlin, Chandran Kukathas, and Fred Schick. I am also grateful for the many useful remarks made at seminars where the article was presented including the *Ethics* symposium in Chicago, where Brian Barry and Karen Cook were commentators. I made use in particular of remarks by Larry Becker, Josh Cohen, Jon Elster, Allan Gibbard, Maggie Gilbert, Ned Hall, Russell Hardin, David Lewis, Steven Lukes, Michael Otsuka, and Michael Smith. The article was finalized while I held a Visiting Fellowship at Corpus Christi College, Oxford, with visiting facilities at Nuffield College, and I am grateful to both institutions for their support.

1. Thus I follow Jon Elster in acknowledging "Norms of Revenge" (see his article in this issue). But I differ from Elster, in so far as he seems to think that norms cannot double in other roles, say as laws. He apparently thinks that a norm is operative only when people actually deliberate or are led to action in a certain way. This is the source of his objection to a rational choice account of norms, for he also thinks that rational choice requires a particular path to action. I differ from him on both points, as will be clear later.

Ethics 100 (July 1990): 725–755

activity, and their contribution to the achievement of public goals that are of benefit to everyone. True, other visions imagine the desired social pattern emerging, as by an invisible hand, from the interactions of individuals who may themselves lack any sense of that outcome, and in these pictures norms play no role in the generation of order. But even those visions, like the ones that give pride of place to norms, assume at least that the order they envisage will not be undermined by the emergence of antisocial norms, say norms of cooperation among manipulative or criminal subgroups.

This inevitable appeal to norms raises the question motivating my article. The question is, What, if anything, makes certain norms resilient; what ensures that in suitable circumstances those norms can be relied on to emerge and persist? In a phrase, what constitutes *virtus normativa?* Unless this question can be answered, then the visionaries who hail certain norms as desirable, for example, can never be sure that the norms are dependable; thus they can never be sure that they are not indulging impractical utopian dreams. The visionaries who are challenged include at one extreme anarchists and at the other those who believe in a strong state, say a state of a social democratic kind. Anarchists have to show that people would behave in a manner conducive to social life without a state to restrain them.[2] Social democrats need to show that the public officials and politicians whom they would empower can be relied on to pursue the public interest.[3]

The article is written within the tradition of rational choice theory. That theory starts from the assumption that two sorts of factor explain a good deal of human behavior. The assumption is almost canonically formulated by John Harsanyi. "People's behavior can be largely explained in terms of two dominant interests: economic gain and social acceptance."[4] The theory suggests that norms will be resilient if—though not necessarily only if—circumstances are such that it is in people's individual interest, economic or social, to honor them. It will be in their economic interest, broadly conceived, if the direct self-interested benefit of honoring the norms, in particular the sort of benefit that can be assigned monetary value, exceeds the cost; it will be in their social interest if honoring the norms promotes the esteem, affection, or pleasure with which they are viewed and this indirect self-interested benefit exceeds the cost.[5] The

2. See Michael Taylor, *Anarchy and Cooperation* (London: Wiley, 1976).

3. See Philip Pettit, "Towards a Social Democratic Theory of the State," *Political Studies* 35 (1987): 42–55.

4. John Harsanyi, "Rational Choice Models of Behavior versus Functionalist and Conformist Theories," *World Politics* 22 (1969): 513–38. The postulate is quoted with approval in Michael Taylor, "Rationality and Revolutionary Collective Action," in *Rationality and Revolution,* ed. Michael Taylor (Cambridge: Cambridge University Press, 1987), p. 66.

5. Here I am suggesting a reading of Harsanyi's postulate under which only current social acceptance matters. What is also undoubtedly important is the social acceptance given to an agent in the past—say, by parents—for certain forms of behavior and the good feeling therefore promoted by such behavior. See John Braithwaite, *Crime, Shame and Reintegration* (Cambridge: Cambridge University Press, 1989).

task of showing whether certain norms are resilient in certain circumstances comes down to that of seeing whether it is possible to derive those norms, under these circumstances, from the assumption that people satisfy Harsanyi's postulate.

Harsanyi's postulate requires some comment even at this early point. As I interpret it, it says that the fact that an option promises to promote an agent's economic gain or social acceptance makes it *pro tanto* desirable. Although the postulate represents gain and acceptance as dominant interests, it does not weight them against other goods or against each other. Thus it enables us to make firm predictions about an individual only for choices where one option does better by gain or acceptance and is not otherwise very costly.[6] In predicting in this way that people will not generally frustrate their economic and social interests, the postulate does not allege that they explicitly calculate about gain and acceptance. The idea is that, whatever the basis on which they make their choices, the fact that a sort of choice which they are in the habit of making becomes inimical to those interests will at least make them pause. No choice of a kind they commonly make is likely to undermine both their economic and social prospects.

This article is not intended as an impartial overview of the different rational choice ways of deriving norms, though an overview is sketched in passing. The main point is to defend a partisan thesis: that the standard mode of derivation adumbrated, if not always spelled out, in the rational choice literature is only one possibility and that we should also pay attention to a sort of derivation which that literature generally derides. The standard strategy of derivation is behavior-based; the strategy identified here is attitude-based. They are not incompatible approaches, and my case for the attitude-based strategy is not meant to cast doubts on the more standard alternative. The intention is to open doors, not to close them. I believe that some norms may only be derivable by the standard strategy, others only by that which I propose, as I believe that some norms may be derivable by both strategies, and some by neither.

The article is in five sections. In the first I offer a definition of norms, elaborating on the observations made above. In the second I distinguish the two strategies of derivation and try to explain why the attitude-based strategy has been ignored. Then in the third section I show how an attitude-based derivation might go. In a short fourth section I look at the definition and derivation of norms, if norms are assumed to require not just fulfillment of the conditions mentioned in the first section but also the common belief that those conditions are fulfilled; this section is something of an appendix to the main paper and may be skipped without serious loss. Finally, in a short conclusion I summarize results so far and show that those who think the rational choice approach cannot make

6. Even without weighting of course, the postulate will enable us to make the prediction that as a sort of option becomes more hostile to either interest, it is less likely to be chosen by a random individual, and it will be chosen less often in the relevant population.

room for the moral aspect of many norms may be mistaken; I sketch an attitude-based derivation for a norm of moralizing about conformity to other norms.

How original is the attitude-based style of derivation that I propose in this article? The derivation will be a novel offering in rational choice circles, as already mentioned; it assumes a rejection of the view, hallowed within those circles, that enforcing a norm necessarily imposes costs on the enforcers. But if the thesis is an original offering in this context, I should stress that it will not appear so in the broader historical picture. The thesis links up with the line about norms which Adam Smith defends in *The Theory of Moral Sentiments.* I might have taken this passage, for example, as my text. "What reward is most proper for promoting the practice of truth, justice and humanity? The confidence, esteem and love of those we live with. Humanity does not desire to be great, but to be beloved."[7]

THE DEFINITION OF NORMS

Almost all accounts of norms emphasize at least these two requirements. First, that if a regularity is a norm in a society, then it must be a regularity with which people generally conform; lip service is not enough on its own.[8] And second, that if a regularity is a norm, then people in the society generally approve of conformity and disapprove of deviance: they may believe, for example, that everyone ought to conform, that conformity is an obligation of some sort.[9]

My inclination is to honor both of these requirements in defining norms. Perhaps the best argument for the first is suggested by the opening sentence of R. M. Hare's *Language of Morals*: "If we were to ask of a person 'What are his moral principles?' the way in which we could be most sure of a true answer would be by studying what he *did.*"[10] If we want to identify a society's norms, then equally the best way is surely by studying what people do there. And that means that a norm is a regularity with which most people in the society must conform. There are regularities which fail this requirement while meeting the second but we shall not cast them as social norms; we might describe them as social standards.

7. Adam Smith, *The Theory of Moral Sentiments,* ed. D. D. Raphael and A. L. Macfie (Indianapolis: Liberty Classics, 1982), p. 166.

8. See David Shwayder, *The Stratification of Behavior* (New York: Humanities Press, 1965), p. 253; Kent Bach and R. M. Harnish, *Linguistic Communication and Speech Acts* (Cambridge, Mass.: MIT Press, 1979), p. 271; Robert Sugden, *The Economics of Rights, Co-operation and Welfare* (Oxford: Basil Blackwell, 1986), p. 166; Robert Axelrod, "An Evolutionary Approach to Norms," *American Political Science Review* 8 (1986): 1097; and Michael Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987), p. 29.

9. See David Lewis, *Convention* (Cambridge, Mass.: Harvard University Press, 1969), p. 97; plus Shwayder; Bach and Harnish; Sugden; Axelrod, "An Evolutionary Approach to Norms"; and Taylor, *The Possibility of Cooperation.*

10. R. M. Hare, *The Language of Morals* (Oxford: Oxford University Press, 1952), p. 1.

The second requirement hardly needs defending, since a regularity would clearly fail to be a norm of a society unless it commanded general commendation in the society.[11] But there is a question about how precisely to define it. In fact there are a number of questions. Ought we to allow the approval or disapproval to be based, say, on the benefit or harm to the agent himself or should we require that it be based on the public interest, at least as the agent sees that? Should we think of approval and disapproval as something which anyone can give anyone else or as something only available in any instance from designated others? Ought we to require that everyone approves of conformity, and disapproves of deviance, in relation to everyone's behavior, his own included, or only in relation to everyone else's? Should we be content if everyone approves or disapproves case by case—*in sensu diviso*—or do we stipulate that everyone has an attitude to the general state of affairs: everyone approves or disapproves *in sensu composito?*[12] Finally, do we really want the requirement to stipulate "everyone all of the time" or is "nearly everyone most of the time" going to be enough?

It would be exceedingly tedious to argue these questions one by one, trying to find the answer which best honors common usage. I propose to identify the right answer in each case by a methodological consideration: that we should make it as easy as possible in this regard for a regularity to count as a norm. The matters raised in these questions are ones to which we commonly pay no attention—we overlook the distinctions in play—and it would be bad practice to define a norm in a way which required that we took a stricter rather than a more relaxed view of those matters. Applying this principle then, and letting "nearly everyone" stand for "nearly everyone most of the time," the second requirement will be this: that nearly everyone, on whatever basis, approves *in sensu diviso* of nearly everyone else he finds conforming—that is, approves of his conforming, approves of him so far as he conforms—and disapproves of nearly everyone else he finds deviating.

Many will object that this lax version of the second requirement employs a notion of approval—and equally disapproval—which is not to be found in everyday usage. The objection is that if I approve of an action only because it suits my particular purposes, then that is not approval, properly speaking; approval proper must be based on principle or must be suited to the social role of the approver.[13] I am not worried about this objection, since I do not care very much about picking up everyday usage. But in any case I think that my generous notion of

11. It might be a norm of course of a subgroup in the society without being generally commended; it would only be required to be commended in the subgroup (see Axelrod, "An Evolutionary Approach to Norms").

12. On this distinction, see Lewis, *Convention,* pp. 64–66.

13. See Philippa Foot, "Approval," in her *Virtues and Vices* (Berkeley: University of California Press, 1978).

approval does have everyday resonance. We speak of someone's expressing approval, not just when he judges an action right or best all things considered but also when he simply likes it. Approval in my sense is nothing less than that broad sort of attitude to which acts of expressing approval testify; it is what expressions of approval express.

But though my lax formulation of this second requirement makes it as easy as possible for a regularity to count as a norm, it does not make it as easy as you might think. Notice in particular that not only must conformity attract approval; equally, deviance must elicit disapproval. Indeed the negative claim is the crucial one, for if we disapprove of someone's not Φ-ing and do not disapprove of his Φ-ing, that is tantamount to approving of the Φ-ing. The claim means that a practice of supererogatory virtue, even one that becomes fairly general, is not going to count as a norm of the society. Why this restriction? Because here we reach a limit where further laxity would put us misleadingly out of line with everyday usage. That a sort of action is normative in a society is not compatible in ordinary parlance with its being regarded as supererogatory. And so I prefer the stricter formulation. The strictness in question may not come to much of course. It is unlikely that a supererogatory type of act will be so commonly performed as to meet the first requirement of general conformity.

Do these first two requirements call for any obvious supplement in the definition of a norm? I believe they do, though the supplement I have in mind is almost universally ignored in the literature. It is surely not going to be enough for normative status that a regularity commands general conformity and that conformity attracts approval, deviance disapproval. For what if there is no connection between these two facts; what if the approval and disapproval are epiphenomenal, playing no part in ensuring the conformity? In such a case I think it is clear that we would hesitate to regard the regularity as a norm. Not that there are any obvious examples of such a case in the offing.[14] It's just that with all the regularities we actually regard as normative, we see the pattern of approval and disapproval as contributing, at least in some way, to the conformity. That is why we lay stress on this pattern in trying to inculcate the norms in our children.

We should add therefore a third requirement to our first two. This is a requirement to the following effect: that the fact that nearly everyone approves appropriately of conformity and disapproves of deviance helps to ensure that nearly everyone conforms. The requirement is not, of course, that people are moved by the consideration that nearly everyone approves and disapproves in the appropriate pattern. It does not matter

14. We do describe rules of logic as norms of reason, and it is sometimes urged that our conformity to these is explicable in evolutionary terms (see Neil Tennant, "Two Problems for Evolutionary Epistemology," *Ratio* 1 [1988]: 47–63). But surely our conformity is not wholly explicable in these terms; surely we are responsive also to the approval which conformity wins and the ridicule which deviance hazards.

what considerations move people deliberatively, what considerations come up in their practical reasoning. At least it does not matter so long as the fact that nearly everyone approves and disapproves appropriately helps to ensure that nearly everyone conforms. That condition would certainly be fulfilled were people to be moved by the consideration of what others approve and disapprove but such reasoning is not required. The condition will be fulfilled, for example, if the considerations that move agents to conform are ones whose relevance or weight is due to the pattern of other people's approval and disapproval; they might be considerations which have been made salient by the approval and disapproval of others, such as considerations as to the goodness and badness of certain options. Equally the condition will be fulfilled if the approval or disapproval of others would come into play and help to produce conformity in the event of a failure by whatever considerations are operative now—say, economic ones—to support such conformity: that is to say, if the approval and disapproval of others serve as standby supports for conformity.[15]

We have identified three requirements that certainly ought to be built into the definition of a norm, two of them commonly recognized, one—perhaps because it is so obviously necessary—not. Recent accounts of norms also tend to build in a further sort of requirement. This is that not only should requirements like those we have canvased be fulfilled, it should also be a matter of common belief that they are fulfilled. I find this requirement congenial, but I propose to ignore it for the moment. We will return in the fourth section below to the case for honoring it in the definition of norms and to the possibility of deriving norms, thus redefined.

The requirements assembled so far are sufficient to give us a workable definition of norms. It goes like this.

> A regularity, R, in the behavior of members of a population, P, when they are agents in a recurrent situation, S, is a *norm* if and only if, in any instance of S among members of P,
>
> 1. nearly everyone conforms to R;
> 2. nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else's deviating; and
> 3. the fact that nearly everyone approves and disapproves on this pattern helps to ensure that nearly everyone conforms.[16]

This definition is modeled on David Lewis's definition of a convention, differing from some versions of that definition only in clauses 2 and 3. Lewis's corresponding clauses are that nearly everyone expects nearly

15. Notice that my formulation of the third requirement allows me to think that a norm may also be a law, even a law such that people's actual reason for conforming to it is the penalty attached. Here there is a contrast with Elster's approach in "Norms of Revenge."

16. If norms are thought to come in degrees, the phrase "if and only if" can be replaced by "to the extent that" (see Axelrod, "An Evolutionary Approach to Norms," p. 1097).

everyone else to conform and nearly everyone prefers to conform on condition that the others do, since universal conformity solves a coordination problem.[17] It is important to recognize, however, that these definitions are in no way exclusive of one another. It is more than likely that a regularity which is a convention in a society will also be in our sense a norm. The point comes up again in the next section.[18]

Beyond my earlier remarks I have little to say in defense of this definition of norms. I believe that the definition catches an interesting category of regularities, even if the category does not fit exactly with everyone's conception of a norm. Thus I hope that even those who question the definition in some manner will still find it a useful way of identifying a topic for discussion. They may not see the topic as involving norms, but the difference need not be more than terminological.

The only thing I will add in defense of the definition is that it includes the sort of regularities that H. L. A. Hart had in mind in his classic discussion of rules of obligation, though it also encompasses more. Hart characterizes such rules by a number of features: they are supported by serious social pressure; they are thought necessary for social life or some prized feature of social life; and they may be individually burdensome, despite being thought to be collectively beneficial.[19] These features are not all mentioned in my definition, but it is a fair bet that anything that has them will satisfy the definition.

TWO STRATEGIES FOR DERIVING NORMS

David Lewis's account of conventions, the model for any work in this area, does more than offer us a definition. It also helps to show why certain regularities can be depended on to emerge and persist as conventions. The key to this aspect of his account is, first, that in a certain sort of coordination predicament it will be rational for each to prefer to follow any one of the regularities possible there if he expects others to follow it; and second, that factors like precedent, salience, and agreement will often identify one regularity as that which others may be expected to follow. Thus it will be rational for each to drive on the left rather than the right if precedent—and perhaps precedent only—means that he expects others to drive on the left. Everyone's driving on the left will emerge as an equilibrium in the sense that no one benefits by unilateral defection from it, and as a coordination equilibrium in the sense that no one benefits by anyone else's unilaterally defecting from it either.

17. See Lewis, *Convention*, p. 42.

18. Margaret Gilbert, "Notes on the Concept of a Social Convention," *New Literary History* 14 (1982–83): 225–51, taxonomizes things so that social conventions are a larger subclass of the class of norms than Lewis would make them.

19. H. L. A. Hart, *The Concept of Law* (Oxford: Oxford University Press, 1961), pp. 84–85. See also Edna Ullmann-Margalit, *The Emergence of Norms* (Oxford: Oxford University Press, 1977), pp. 12–13.

Lewis does not say that rational calculation in such a case is what makes each conform to the regularity; the immediate trigger may be the training received, a habit ingrained by the training, even a compulsion to be conformist. His claim is best taken as follows: that so far as it is rational for each to conform to any regularity that constitutes a convention, that makes it very probable that he will conform. He may actually conform from habit, but the rationality of conforming makes it likely that even if the habit disappeared, the conformity would tend to continue, if only after a lapse.[20] The rationality of conforming programs for the resilience of the conformity, even if it does not produce the conformity; it more or less ensures that whatever productive mechanism generates the agent's behavior—habit, rule of thumb, calculation—it will generate behavior in conformity to the convention.[21]

Our definition of norms does not serve on its own, unlike Lewis's definition of conventions, to show that certain regularities can be depended on in certain conditions to constitute norms. Such a derivation would provide us with an understanding of why certain norms emerge and/or persist and perhaps why other norms fail to do so. It might not shed light on the precise process of emergence or persistence—here socialization is probably the most important factor—but it would do something as good or better. It would show why in certain conditions those norms more or less had to emerge or more or less have to persist. The challenge then is to supplement our definition of norms with a derivation, or at least a derivation for some of the norms defined: a story as to why those norms can be depended upon to emerge and persist under certain circumstances.

Looking at our definition of norms, two strategies of derivation suggest themselves. One strategy would be to show first why certain behavioral patterns are intelligible and then to explain why, having appeared, they should attract the sort of approval that constitutes them as norms. The other would take the contrary path, explaining first why certain attitudes of approval are intelligible and then showing how they might generate the patterns of behavior required for norms. The first strategy is behavior-based, the second is attitude-based. In terms of Harsanyi's postulate, the first would tend to show that the behavior is economically rational and, being performed by nearly all, comes to be socially

20. Lewis makes a complementary point: "If that habit ever ceased to serve the agent's desires according to his beliefs, it would at once be overridden and corrected by conscious reasoning" (David Lewis, "Languages and Language" [1972] in his *Philosophical Papers*, vol. 1 [Oxford: Oxford University Press, 1983], p. 181). On related matters, see Philip Pettit and Michael Smith, "Backgrounding Desire," *Philosophical Review* (in press).

21. On this notion of programming, see Frank Jackson and Philip Pettit, "Functionalism and Broad Content," *Mind* 97 (1988): 381–400, "Program Explanation: A General Perspective," *Analysis*, vol. 50 (1990), and "Structural Explanation in Social Theory," in *Reductionism and Anti-reductionism*, ed. D. Charles and K. Lennon (Oxford: Oxford University Press, in press).

rational too; the second would show that it is socially rational from the start.

David Lewis indicates how we might pursue the behavioral strategy in arguing that conventions, once established, are likely to constitute norms; as well as the first, they are likely to meet the second and third clauses in our definition of norms. Lewis's conclusion, in his own words, is that "one is expected to conform, and failure to conform tends to evoke unfavorable responses from others. . . . These are bad consequences, and my interest in avoiding them strengthens my conditional preference for conforming."[22] Without going into the detail of his argument, we may note that it turns crucially on propositions like these.

> 1. Universal conformity with a convention like driving on the left is a coordination equilibrium in the sense that not only does no one benefit by unilaterally defecting himself, equally no one benefits by anyone else's unilaterally defecting either: in fact everyone is usually made worse off by anyone's unilateral defection.[23]
>
> 2. Everyone therefore will tend to disapprove of anyone else's defecting from such an outcome, so that the second condition in our definition of norms will be effectively fulfilled.
>
> 3. Since everyone is in a position to realize this, everyone has an extra motive not to defect from the outcome, over and beyond the fact that it would bring him no benefit: namely, that he would thereby attract the disapproval of others. Thus the third condition in our definition of norms will also be fulfilled.

The rational choice literature of the past decade or so supports the behavioral strategy for deriving norms in two ways. First of all, it makes explanatory claims sufficient to support such a strategy.[24] And second it presents an argument against the alternative attitude-based approach. In this section I will look at those explanatory claims, suggesting that in some ways they may be overblown, and I will show that the argument against the attitude-based strategy is almost certainly misconceived. Thus I prepare the way for the attitudinal derivation of certain norms explored in the next section. I should stress again that there is no need to reject one strategy of derivation because of recognizing the other. I think that the attitude-based strategy deserves more attention than it has received, but I do not hold that it is uniquely right. Some norms may be derivable in the one way, some in the other; some norms may be subject to both sorts of derivation, as some will be subject to none.

The norms that have been at the focus of concern in the rational choice literature are those such that conformity to them enables people to resolve free-rider problems, in particular problems that are also many-

22. Lewis, *Convention,* pp. 99–100.

23. The equilibrium, in Lewis's terminology (*Convention*), is a proper equilibrium, so far as everyone does worse by unilaterally defecting. Such a proper equilibrium is an instance of what Sugden defines as a stable equilibrium (p. 28).

24. Sugden explicitly develops a behavior-based strategy of derivation.

party prisoner's dilemmas.[25] In a prisoner's dilemma each party faces options of cooperating or defecting in some way and the following two conditions are fulfilled: universal cooperation is Pareto-superior to universal defection, being better for some—perhaps for all—and worse for none; but defecting is the dominant option, being better for each regardless of what others do. Arguably, conforming to norms like the following is equivalent to cooperating in a many-party prisoner's dilemma, so that universal conformity—though, in most cases, just fairly general conformity will do—represents an escape from the predicament.

1. Telling the truth reliably rather than expediently, randomly, or whatever.
2. Keeping promises reliably.
3. Refraining reliably from theft or fraud or violence.
4. Reliably discharging any publicly assigned duties.
5. In general, reliably contributing to goals that are of benefit to everyone.

How might we explain the emergence and persistence of behavior in accordance with such norms, abstracting for the moment from how the behavior comes to attract approval? That universal behavior of the kind in question would enable people to resolve prisoner's dilemmas does not itself furnish an explanation of emergence and persistence, though some authors write, misleadingly, as if it did. Thus Edna Ullmann-Margalit writes, "Such situations 'call for' norms. It can further be said that a norm solving the problem inherent in a situation of this type is generated by it."[26] In an individual prisoner's dilemma all do better if all conform to a normative resolution than if all defect, but each does better still if he defects while the others conform. So why should universal conformity emerge or persist?

One now standard answer is motivated by the observation that the parties who conform, if they do conform, face an indefinitely extended sequence of prisoner's dilemmas, not a single one, and that in such a sequence permanent defection is not a dominant option; it is not the best for each regardless of what others do. Permanent defection by all may be an equilibrium outcome, in the sense that no one can unilaterally depart from it with benefit. Equally permanent conformity or cooperation by all may not be an equilibrium outcome. But, as Michael Taylor has shown, there are equilibrium outcomes besides permanent defection by all, at least under plausible assumptions such as that people do not severely discount future benefits. And some of the other outcomes are Pareto-

25. This trend is breaking down (see Sugden; and Taylor, *The Possibility of Cooperation*). On the relation between free-rider problems and prisoner's dilemmas, see my paper, "Free Riding and Foul Dealing," *Journal of Philosophy* 83 (1986): 361–79, reprinted in *The Philosopher's Annual* 9 (1986): 149–67.

26. Ullmann-Margalit, p. 22. For a discussion of other such views, see Anthony Heath, *Rational Choice and Social Exchange* (Cambridge: Cambridge University Press, 1976), chap. 7.

superior to permanent defection by all.[27] The most salient example of such an outcome is that under which each tit-for-tats in some way: he begins by cooperating but only cooperates in a later round if no one defected (nonpunitively) in the previous round. This is an equilibrium, because anyone who unilaterally defects will be punished by the defection of others and will have to cooperate while they defect (in punishment for a previous defection of his) before they return to cooperation; thus any one who unilaterally defects will suffer through doing so. The outcome of universal tit-for-tat is Pareto-superior to that of permanent defection by all because it means that everyone is better off, benefiting from universal cooperation rather than universal defection at each round.

The fact that joint tit-for-tat is an equilibrium outcome which is Pareto-superior to permanent defection by all suggests an explanation for why universal tit-for-tat behavior should emerge and persist. It will emerge if each can persuade others that he is a tit-for-tatter, so that it is to their advantage to tit-for-tat with him. It will persist if each recognizes that a unilateral defection will attract the punitive defection of others, so that he does better continuing to tit-for-tat and therefore, assuming that others tit-for-tat with him, continuing to conform.

If this explains why people might evince tit-for-tat behavior, what might explain the approval for that behavior which is required if it is to constitute a norm? Here the crucial fact is not that universal tit-for-tat is an equilibrium but, as Russell Hardin has emphasized, that it is also a coordination equilibrium.[28] It is an outcome such that each is made worse off by anyone else's unilateral defection, since each is forced to defect at the next round in punishment, thereby jeopardizing the benefits of general cooperation. Hence it is a coordination equilibrium: no one benefits—in fact each suffers—by anyone's unilateral defection, his own or someone else's. That being the case, we can invoke propositions like those mentioned in discussing Lewis's argument that conventions are likely to be norms, in order to explain why people may be expected to disapprove of anyone else's unilaterally defecting, giving everyone extra reason not to defect himself.[29] People will disapprove of anyone else's unilateral defection, since any such defection harms each of them.[30] That is a fact which everyone is in a position to recognize and, since disapproval

27. See Taylor, *Anarchy and Cooperation*, and *The Possibility of Cooperation*.

28. Russell Hardin, *Collective Action* (Baltimore: Johns Hopkins University Press, 1982), p. 171.

29. The point is not generally recognized. It would have helped Russell Hardin himself at pp. 105–6 of *Morality within the Limits of Reason* (Chicago: University of Chicago Press, 1988), as it would those theorists who rely on the adage that the customary becomes obligatory; they are discussed in Heath, pp. 65–67, 161–62. One writer, however, who defends a similar claim is Sugden, p. 166.

30. If this claim seems questionable, see the discussion in the next section. Notice in particular that disapproval is an attitude: a disposition to express disapproval, if the circumstances are suitable.

is generally a bad, the recognition will give everyone an extra motive not to defect. Hence it appears that any tit-for-tat regularity will constitute a norm. Not only will it attract general conformity. Everyone will approve of anyone else's conforming, at least to the extent of disapproving of anyone else's unilateral defection, and this will help to ensure that there is indeed general conformity.[31]

We have sketched a behavior-based derivation, not of a norm like that of reliably telling the truth or keeping promises, but of a closely related norm: that of truth-telling or promise-keeping in a tit-for-tat way. The derivation is of interest because if everyone tit-for-tats in truth-telling everyone will behave as he would do were he telling the truth reliably: it will be as if everyone were telling the truth reliably. The derivation works, notice, on lines parallel to those explored by Lewis. In the Lewis case, agents have to identify a regularity on which to coordinate among a set of equally attractive conventions: say, driving on the left or the right. In this case things are set up so that they have the parallel problem of coordinating on one of those regularities that yield a superior equilibrium outcome to permanent defection. They have to coordinate on strict tit-for-tat, for example, or on any of the equally attractive variations: say, tit-for-double-tat, tit-for-tat-by-a-certain-number, and so on.

This is sufficient to show that the explanatory claims of recent rational choice theory serve to underpin a behavior-based derivation of certain norms. How successful that derivation is depends on how plausible those claims are. It is not a part of my brief to undermine such claims, but I would like here to mention two reservations about the tit-for-tat derivation. A first is this. A rational-choice derivation need not posit rational calculation—we saw this in discussing Lewis—and a tit-for-tat derivation need not therefore impute tit-for-tat reasoning. But a tit-for-tat derivation predicts that people will break norms punitively, in order to punish those who break them for convenience, even if the punishment is not explicitly rationalized in tit-for-tat terms. And this disposition is not generally manifested among those who honor norms; it is present, at most, only in certain sorts of cases.

My second reservation stems from a distinction, on which I have written elsewhere, between type A and type B prisoner's dilemmas.[32] In a type B dilemma, defection by even a single individual plunges at least one cooperator, and perhaps many more, below the baseline of universal defection. In a type A dilemma this is not so and, at the limit, the lone defector may have only an imperceptible negative effect on cooperators: the effect, say, of the one remaining person who continues to use chlorofluorocarbon sprays. My reservation about tit-for-tat derivations is that

31. Notice of course that under this derivation everyone will approve of defecting—and disapprove of conforming—when the defection is a tit-for-tat punishment.

32. See Philip Pettit, "Free Riding and Foul Dealing," and "Foul Dealing and an Assurance Problem," *Australasian Journal of Philosophy* 67 (1989): 341–44.

in a type A dilemma, it is not clear that anyone will be able to make it credible to the potential free rider that he is a tit-for-tatter. In particular it is not clear that he will be able to make credible the threat to defect—and put at risk all that has been achieved—just in order to punish a lone, barely irritating defector. The answer to this may be to accept that the only norms which can be derived under the tit-for-tat approach resolve type B dilemmas: roughly, what I have called "foul dealer" as distinct from free rider problems. But it seems clear that that would be a substantial concession.[33]

We have seen that the explanatory claims of recent rational-choice theory are naturally deployed to support a behavior-based derivation of norms, in particular a derivation of norms other than just the conventions covered in Lewis's treatment. That may be one reason why rational-choice theorists have not given much thought to the possibility of an attitude-based derivation. But there is a second reason that has certainly been of importance in directing attention away from such a derivation. This is that, within rational-choice theory, it has become established wisdom that any attitude-based approach falls foul of a decisive objection.

An attitude-based derivation of norms would try to show that a certain sort of behavior is bound to attract approval, its absence disapproval, and that such sanctions ought to elicit the behavior required, thus establishing norms. The objection is that any derivation of this kind supposes, illicitly, that the enforcement of norms—the sanctioning of conformity and deviance—is costless and will be happily conducted by people in general. James Buchanan puts the opposite, standard view. "Enforcement has two components. First, violations must be discovered and violators identified. Second, punishment must be imposed on violators. Both components involve costs."[34]

The objection in play is often developed in the form of a paradox. Norms may often serve to get us out of collective action predicaments like the prisoner's dilemma: they elicit a sort of action such that everyone is better off if everyone adopts it, and they do this even when each is

33. Notice, however, that the tit-for-tat story considered here is only one of a number of related accounts (see Taylor, *Anarchy and Cooperation*). One account of particular interest would explain behavior like general truth-telling or promise-keeping as the outcome, not of tit-for-tatting in a single many-party dilemma, but of tit-for-tatting in various two-party dilemmas (see R. Hardin, *Morality within the Limits of Reason*, p. 105). Such a possibility should not surprise us, since two-party dilemmas are by definition of type B: the lone defector makes the cooperator worse off than he would be under joint defection. On tit-for-tat in two-party dilemmas, see Robert Axelrod, *The Evolution of Cooperation* (New York: Basic, 1984). On how a tit-for-tat type of strategy may even be rational in a sequence of such dilemmas of known finite length, see Philip Pettit and Robert Sugden, "The Backward Induction Paradox," *Journal of Philosophy* 86 (1989): 169–83; and Christina Bicchieri, "Self-refuting Theories of Strategic Interaction," *Erkenntnis* 30 (1989): 69–80.

34. James Buchanan, *The Limits of Liberty* (Chicago: University of Chicago Press, 1975), pp. 132–33. See too Heath, pp. 156–58; and Axelrod, "An Evolutionary Approach to Norms," p. 1098.

motivated to choose a different option. But the objection suggests that norms can persist only if we find some other way of escaping a similar predicament which is raised by their enforcement. Everyone is better off if everyone enforces a norm, but because enforcement is costly each is motivated not to bother enforcing it himself. And so norms can solve certain collective action predicaments only if the collective predicaments they in turn generate can be solved by something else.

Anthony Heath makes the point in connection with a norm of output-restriction among a large group of workers. "Enforcement of the norm is assuredly a public good: I will get the benefits whether or not I actually do the enforcing and will hence prefer to leave the embarrassing task of disciplining the rate-busters to others. So will everybody else. And so the rate-busters will go unchecked."[35] Michael Taylor makes the point more generally: "The maintenance of a system of sanctions itself constitutes or presupposes the solution of another collective action problem. Punishing someone who does not conform to a norm—punishing someone for being a free rider on the efforts of others to provide a public good, for example—is itself a public good for the group in question, and everyone would prefer others to do this unpleasant job. Thus, the 'solution' of collective action problems by norms presupposes the prior or concurrent solution of another collective action problem."[36]

This line of objection, common though it is, rests on a mistake. It assumes that the enforcement of norms must involve intentional action and since action always generates at least time costs that it must therefore be potentially costly for those who conduct it. The surprising thing however is that this is false. Buchanan mentions two sorts of enforcement costs: those of identifying violators and those of disciplining them. But people do not have to identify violators intentionally; they just have to be around in sufficient numbers to make it likely that violators will be noticed. And equally, people do not have to discipline violators intentionally, going out of their way for example to rebuke them or report them to others;[37] they just have to disapprove of them—or at least be assumed to disapprove of them—whether that attitude ever issues in intentional activity.

It will be readily conceded that given sufficient numbers, enforcement need not involve intentionally seeking out the violators of a norm. What will come as a shock to many, however, is the claim that a violator can be punished—or of course a conformer rewarded—by the attitudes of others, even when those attitudes are not intentionally expressed, say,

35. Heath, p. 158.

36. Taylor, *The Possibility of Cooperation*, p. 30.

37. Braithwaite has suggested (*Crime, Shame and Reintegration*) that in any case reporting violators to others is generally something people enjoy and that the argument may also break down here. In order for the suggestion to work, of course, people must not enjoy falsely reporting violations nearly as much as doing so truthfully. On related matters, see the essay on gossip in John Sabini and Murray Silver, *Moralities of Everyday Life* (Oxford: Oxford University Press, 1982).

in censure or praise. Yet the point, once put, is fairly obvious. We care not just about the rebukes and commendations we receive from others but also about whether they take a negative or positive view of what we do: look at the eagerness with which we search for cues as to the view they actually take. We care about their dispositions to rebuke or commend us, even if the costs—say, the costs of social embarrassment—mean that those dispositions are not much exercised. How can we know about other people's dispositions if they are not exercised? Easy. We know what they know of us and, ascribing similar standards to them, we know whether they are likely to think well or badly, to take a favorable or unfavorable attitude.[38]

But not only is it fairly obvious that even in the absence of praise or censure the attitude of approval is a good that I can savor and the attitude of disapproval a bad under which I may smart; the claim is also supported by tradition. As with many other propositions in this article, Adam Smith can be invoked as a relevant authority: "We are pleased to think that we have rendered ourselves the natural objects of approbation, though no approbation should ever actually be bestowed upon us: and we are mortified to reflect that we have justly merited the blame of those we live with, though that sentiment should never actually be exerted against us."[39]

The rational choice tradition has been blind to the fact that the goods which we seek from others include goods that they do not intentionally bestow, in particular attitude-dependent goods like approval and disapproval. One reason may be that in giving us the distinction between strategic and parametric rationality, rationality exercised respectively with and without the assumption of rationality in the causally relevant environment, the tradition naturally suggests that parametric rationality is suited for dealings with nature, strategic for dealings with other people. Even though he doesn't endorse the suggestion fully, the distinction leads Jon Elster, for example, to the following view. "Strategic rationality is defined by an axiom of symmetry: the agent acts in an environment of other actors, none of whom can be assumed to be less rational or sophisticated than he is himself."[40]

38. If further explanation is needed, then one way of explaining why we care about covert as well as overt approval is that someone's covert attitudes affect how he will later speak of us and deal with us. This explanation may be given a sociobiological gloss, accounting for why we care even about the views of those we may never knowingly meet again: for example, the pedestrian who sees me driving through a red light and clearly regards me in a negative way. More on this in the next section.

39. A. Smith, p. 116. Here and elsewhere Smith wants to be able to say that we desire not only to be such that others are disposed to praise us but also to be such that others are rightly disposed to praise us. I suspect that he illicitly uses the first claim to make the second, intuitively stronger, thesis seem plausible. See also p. 310 where he makes three distinctions when four are obviously being offered.

40. Jon Elster, *Explaining Technical Change* (Cambridge: Cambridge University Press, 1983), p. 77. Notice that Elster, in going on to taxonomize social interdependencies, fails

If strategic rationality is thought uniquely suitable for dealings with other actors, in particular other actors who know as much as the agent knows, then the assumption is that any goods which one agent can seek from others are goods which the others rationally and therefore intentionally bestow. It means, as is indeed often explicitly maintained, that if one agent acts rationally with a view to securing such goods from others, then what he is trying to engineer is a rational exchange. But this emphasis on exchange is not always appropriate. The benefit which an agent seeks from certain others may be a benefit involuntarily provided, as when he gets them to think well of him or at least not to think ill. There need be no element of exchange in the interaction. Thus people can be more or less involuntary enforcers of norms, automatically providing suitable rewards and punishments for acts of conformity and deviance. Buchanan, thinking of electric fences and gun traps, says this: "We need not reach into the extremities of science fiction to think of devices that could serve as automatically programmed enforcers."[41] We may readily agree, for we can imagine ourselves as enforcers of that kind.

In conclusion, I would like to add a thought to bolster the point. Reflecting on the automatic way in which we sanction one another's actions by approving and disapproving, you may well think that what the rational self-interested agent should do is take over this sanctioning in an intentional way and try to drive a harder bargain for the goods he offers or the bads he reserves. But here we confront an extremely interesting and indeed pervasive paradox. When I elicit someone else's approval for an action, without intentional action on that person's own part, I enjoy a good which would not be in the offing were I to realize that the approval was provided intentionally, or at least was provided intentionally on grounds other than that it is deserved. The good of having someone else's esteem or gratitude for an action, even the good of just having him look on the action with pleasure, is something that that person therefore cannot intentionally use in exchange. If it is not enough for him to approve that he understand the merits or attractions of what I have done, if he approves only because he has an extraintentional reason for doing so, or only in part because of this, then the approval loses its significance and value. The point will be familiar. You cannot sell your approval any more than you can sell your friendship or love or trust.[42]

---

to notice the possibility that the action of each may depend on the preference structures—the attitudes—of all: this is the possibility exploited in the argument of section 3.

41. Buchanan, p. 131.

42. See Jon Elster, *Sour Grapes* (Cambridge: Cambridge University Press, 1982), chap. 2, on "essential by-products." On similar points, see Philip Pettit and Geoffrey Brennan, "Restrictive Consequentialism," *Australasian Journal of Philosophy* 64 (1986): 438–55; Philip Pettit, "The Consequentialist Can Recognise Rights," *Philosophical Quarterly* 35 (1988): 537–51, and "The Paradox of Loyalty," *American Philosophical Quarterly* 25 (1988): 163–71.

AN ATTITUDE-BASED DERIVATION

Many norms may lend themselves to a behavior-based derivation. In other words considerations to do primarily with economically rational behavior may explain why certain norms are resilient: why they can be relied on to emerge and persist in certain conditions. But I suspect that the set of derivable norms is larger than the set of norms derivable in that way, and in this section I look at the possibility that certain norms may lend themselves to an attitude-based derivation. The set of norms derivable in this way may overlap with the other set, but it certainly extends beyond it.

The norms of particular interest in moral and political theory are those which would enable people to solve collective action problems, whether problems that arise for the society at large or for particular subgroups. Such problems arise when it appears, usually in the light of considerations of economic gain, that if agents are individually rational then they will generate a Pareto-inferior member of the set of possible outcomes.[43] To solve such a problem is to succeed in getting a Pareto-optimal outcome instead: an outcome which is not Pareto-inferior to any other, there being no other that is preferred by some and that is not preferred by none. In looking at the behavior-based strategy we concentrated, as is usual, on norms that solve one species of collective action problems, namely, prisoner's dilemmas. Here we shall maintain that focus, since it has the virtue of being familiar and our aim is only to see how the attitude-based derivation might go, not to provide a survey of all the norms that are so derivable. It should be remembered, however, that there may well be norms that solve no collective action problems at all, even for a relevant subgroup—say, norms of revenge—and that among those that solve such problems there are certainly norms that solve problems other than prisoner's dilemmas. Conventional norms are of this kind, since even coordination predicaments count as collective action problems: under conditions of ignorance, rational action can lead to a Pareto-inferior outcome.[44]

The key to the attitude-based strategy of derivation is the recognition that there is a cost-benefit structure operative in social life which rational choice theory has generally neglected: the structure associated with people's thinking ill or well of an agent—or being thought to think ill or well—whether they actually censure or praise. I hypothesize that once these approbative costs and benefits are put into the equations, then we can see our way to explaining why the emergence and persistence of otherwise puzzling norms may be unsurprising. In order to support that hypothesis I shall set out a number of fairly plausible assumptions and argue that given those assumptions we should expect the approbative costs and benefits to encourage the emergence and persistence of certain norms.

43. See Taylor, *The Possibility of Cooperation,* p. 19.
44. Here I break with Taylor, *The Possibility of Cooperation,* p. 30.

In effect I shall argue that in conditions where those assumptions are satisfied the norms in question are derivable in an attitude-based way. There are five assumptions in all. I first present the assumptions, indicating briefly why I think that each is plausible. Then I show why we may expect to find certain norms in operation wherever they are satisfied.

The Interaction Assumption

Assumption 1 is that in all human societies there are collective action predicaments with these characteristics. First, among the options available to any agent in the sort of situation involved nearly everyone is better off if everyone else takes one particular option than if everyone else rejects it: the option in question is, in that sense, a collectively beneficial one. Second, and more strongly, everyone is made better off in at least one respect—and better off therefore in most cases, I shall assume—by anyone else's taking the collectively beneficial option: either that person increases or ensures the collective benefit being offered or he makes it more likely that the collective benefit will be or remain an offer.[45] The second condition is stronger than the first because it rules out the possibility that the absolute best result for everyone is not that everybody else takes the option in question but that a certain percentage do so.[46]

This first assumption is satisfied in a variety of interactions, most importantly in various prisoner's dilemmas. If everyone else tells the truth reliably, everyone is better off than if everyone else does so randomly and in one respect everyone is made better off by anyone else's being a reliable truth-teller; he benefits at least indirectly, so far as that person's conformity to the truth-telling norm reinforces truth-telling overall. The case is similar for reliably keeping promises and similar for revealing the nature of your wares, refraining from violence to others, and generally adopting a nonmalevolent stance. Again everyone is better off if everyone else contributes to the provision of nonexcludable goods like a quiet neighborhood or a clean environment than if no one does so and in the relevant respect everyone is made better off by anyone else's contributing to such a good. The examples being offered are familiar, and they need not be further elaborated to vindicate our initial assumption. Note that we have only mentioned examples of predicaments involving the society as a whole. There are bound to be analogous situations for subgroups in any society, but we will not consider them here.

The Publicity Assumption

Assumption 2 is that in at least many of the sorts of predicament described some people will be in a position to know, or be in a position where they

45. This might be weakened, so that what is required is at least that everyone is not made worse off in most cases. The weakening will not affect the argument, provided assumption 4 is strengthened so as to compensate.

46. For complicated possibilities of this kind, see Thomas Schelling, *Micromotives and Macrobehavior* (New York: Norton, 1978), esp. chap. 7.

are likely to come to know, of anyone who acts in a way that promotes the collective benefit that he does so and of anyone who fails to act in that way that he fails. This is an assumption of exposure or publicity. Clearly it is not always satisfied, since there are many occasions when we can fail to do our collective bit and successfully cover our tracks. We can litter the park at night or supply defective goods under cover and so reasonably hope to get away without having the offense put down to us. But equally clear is that the assurance of being able to keep an offense hidden from the eyes or ears of our compatriots is only rarely available. If we choose to offend then in most cases we do so at our own risk.

The Perception Assumption

Assumption 3 spells out something that is implicit on one reading of the last assumption. This is that nearly everyone who knows of someone that he has done or failed to do his collective bit in some way will perceive that that person has acted in a way that is collectively beneficial or nonbeneficial and indeed in a way that is in at least one respect beneficial or nonbeneficial to him in particular. Not only does he know that the person has told the truth, he also knows that this is a sort of action such that everyone is better off if everyone else reliably does it. And he knows that he in particular is better off in one respect for the other person's doing it. The other person's telling the truth may benefit him directly but at least it will benefit him by increasing or making more secure the sort of good he would enjoy through everyone else's telling the truth. Again, to take another example, not only does he know that another person has littered the park, he also knows that this is a collectively nonbeneficial action and he knows that he in particular suffers in some measure from it: his environment is not as clean as it would be if no one else was a litterbug. Such examples should make it clear both that the perception assumption is distinct from the publicity assumption and that in many cases it is equally uncontroversial.

The Sanction Assumption

Assumption 4 is that nearly everyone approves of nearly everyone who benefits him in some respect through performing a collectively beneficial action and disapproves of nearly everyone who harms him through performing a collectively nonbeneficial action. To approve or disapprove in the broad sense adopted here is to be disposed respectively to encourage or discourage the agent in question. That the action is personally beneficial or harmful certainly provides a ground for approval or disapproval: only saints could fail to give it weight. And that the action is collectively at the same time as personally beneficial, collectively at the same time as personally harmful, means that even saints can indulge themselves. They may count their personal gain or loss into their reasoning and, even if that is totally uncongenial—even if they are perfect altruists—they may be moved by the consideration of collective benefit and harm to match the rest of us in our postures of approval and disapproval.

This assumption will not be satisfied in every case. If the beneficial action is very burdensome, for example, then while it may attract approval, many people will not disapprove of the harmful alternative; they will see it as natural and understandable. But it is surely plausible to think that the assumption will be satisfied for at least some of the collective benefits and harms invoked in the interaction assumption. Remember in this connection that while approval and disapproval require an appropriate disposition to encourage or discourage, the property required can be extremely weak. It may be the disposition to do those things only in circumstances where there are no costs whatever involved: that is, in circumstances of a kind unlikely to arise, where the acts in question are not found embarrassing or judgmental, for example, and they do not cost any time or effort that could be better spent. I smart under the gaze of the most uncensorious of my fellows if I realize that, while he will never rebuke me, he would do so were he less unassuming or were social life more conducive to such activities.

The Motivation Assumption

The last of my five assumptions is that people are moved in great part, though not exclusively, by a concern that others not think badly of them and, if possible, that they think well of them. They may not calculate by explicit reference to the opinions of others but what opinions they ascribe will affect what considerations they find salient in deliberation or what considerations they would find salient if the operative considerations supported actions that are offensive to others. This assumption is intuitively acceptable, as we have already emphasized, fitting for example into that long tradition of European thought in which the love of esteem, affection, and acceptance in general is hailed as one of the great human passions.[47] Adam Smith gives forceful expression to the assumption: "Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive."[48]

This assumption should also recommend itself nowadays, for it is built into at least two major schools of contemporary social theory. It is part of the theory of rational choice, as appears from the emphasis in Harsanyi's postulate on social acceptance. And it should also be congenial to those in the sociological tradition of theory. Within that tradition the desire for status ranks with the desire for wealth and power as one of the basic human motives and to enjoy status is to enjoy a special kind of acceptance: specifically, a greater acceptance than relevant others.

47. See, e.g., Arthur O. Lovejoy, *Reflections on Human Nature* (Baltimore: Johns Hopkins Press, 1961), lecture 5.

48. A. Smith, p. 116.

For those who are less impressed than I am with the plausibility, and the traditional endorsement, of the motivation assumption, it may be useful to point to an instrumental reason why people should care about what others think of them, even others who do not say or do anything by way of face-to-face rebuke or punishment. That someone comes to think ill of me for having done something gives me reason to believe that, even if no immediate penalty is forthcoming—the costs are too high to that person—still, the person is thereby made more likely, if the costs are right, to speak unfavorably of me to others and damage my prospects of being favorably treated at their hands, or to damage my prospects directly by treating me unfavorably himself: say, by preferring another in the exercise of some patronage. This is to say that someone's thinking ill of me represents an increased probability of my being ill-treated, so that no one should be surprised that we care about what others think of our actions, even when those others say or do nothing in immediate censure. I actually believe, and I assume here—though not a lot depends on the difference—that what others think is a matter of intrinsic and not just instrumental concern to most of us. Otherwise it is hard to see why we worry, as we surely do, about being noticed doing compromising or demeaning things by complete strangers: say, being noticed peeping through a hotel keyhole, running a red light, or just picking your nose. It may be that this intrinsic concern for what others think is implanted in us for instrumental reasons—reasons that may not themselves make any impact on us—by evolution or by training.

It may be said against the motivation assumption that we only care about acceptance when it is given for reasons of a certain kind or by people in a certain category. The saint does not care for the knave's acceptance, the sadist does not care for the victim's. But this objection is misleading. The saint is put off by the cost of the knave's approval, the sadist by the cost of the victim's: in the one case wrongdoing, in the other kindness. The assumption regains plausibility when we realize that it postulates a desire for the property of being accepted, not a desire for every prospect that involves acceptance. Would the saint or sadist like to continue to act as he does and now in addition have the acceptance of the relevant party? That is the question to be asked and the motivation assumption, plausibly enough, says that, impossible though it might be, the saint and the sadist would each prefer that alternative.[49]

With the five assumptions in place, I am now in a position to argue that under conditions where those assumptions are satisfied, certain norms can be depended on to emerge and persist.

Stage 1. The interaction, publicity, and perception assumptions mean that in any society there will be certain action types that satisfy the conditions

49. On the distinction between desiring properties and prospects, see Philip Pettit, "Decision Theory and Folk Psychology," in *Essays on the Foundations of Decision Theory,* ed. Michael Bacharach and Susan Hurley (Oxford: Blackwell, in press).

they lay down. The action types will be collectively beneficial options such that everyone is better off in some measure for any one else choosing one, worse off for any one else choosing something different. They will be options which no one can choose or reject without someone else being likely to notice. And they will be options such that anyone who notices will recognize the collective and personal benefit or harm occasioned by such a choice.

Stage 2. By the sanction assumption, many of these action types will be such that the choice of an action type will usually attract approval, the choice of an alternative disapproval. Thus the second condition in our definition of a norm will be fulfilled by those action types. Each will be a regularity such that nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else's deviating.

Stage 3. By the motivation assumption, this approval and disapproval constitutes a potential motive for people generally to evince such actions. It seems reasonable to take it that at least if people were not to evince the action types in question, then the motive would become actual: people would become aware of the approval they had lost, the disapproval they had attracted, and this awareness would generate a corresponding concern. Assume for the moment that if the potential motive were to become actual in this way, then that motive would also be generally effective, eliciting the appropriate action types; we redeem this assumption in stage 4, below. It will follow in that case that the existence of the pattern of approval and disapproval in question makes it relatively certain that the first and third conditions in our analysis of a norm will be fulfilled for those action types. Each type will be a regularity to which nearly everyone conforms: either he will have reasons to conform independently of the approval and disapproval or, lacking such reasons and tending not to conform, he will be brought into line by the consequent loss of approval, the consequent attraction of disapproval. And each will be a regularity such that people's conformity to it is more or less ensured—if not actually produced in every case—by the approval given to conformity, the disapproval given to deviance.

Stage 4. The action types will constitute norms, therefore, in any circumstances where the motive in question—the desire to have the approval of others and, in particular, to avoid their disapproval—can be expected to outweigh competing motives, including motives related to the immediate costs of conformity, the threats of powerful agents, the operation of other norms, the feelings of guilt derived from past patterns of approval, or whatever. Among many-party prisoner's dilemmas, we cannot expect the motive to triumph in foul dealer problems, for example, since cooperating exposes each to the risk of being plunged beneath the baseline of universal defection, even by a lone defector. Indeed cooperating may be so burdensome in such a predicament that defecting does not attract disapproval, so that the derivation fails at stage 2. Other things being equal, however, we can perhaps be more sanguine with prisoner's

dilemmas of the other type: for example with free rider problems in which cooperation certainly costs something but at least does not involve the foul dealer risk; so long as a certain minimum of others cooperate too, the cooperator is better off than under universal defection.[50] It would seem that with many action types that represent cooperative options in such predicaments, if the action types satisfy all the conditions mentioned earlier, then they are likely to emerge and persist as norms of the society.

An example will breathe life into this abstract derivation. A particularly appropriate example, given Garrett Hardin's famous analysis of the tragedy of the commons, is a norm of not overgrazing such shared land. According to Hardin we ought to expect a commons to be overgrazed, so far as overgrazing is a dominant option for each: better if others choose it, better if others do not choose it. The tragedy is that overgrazing by all is worse for all than refraining; the situation is a many-party prisoner's dilemma.[51] It turns out, however, that the case is one where our assumptions would lead us to expect a norm of not overgrazing to emerge and persist. Such a norm may explain why, as a matter of fact, the commons system was generally very successful in medieval Europe.[52]

Under commons conditions we have a collective action predicament in which not overgrazing is collectively beneficial and in which each is benefited in some measure by anyone else's not overgrazing: thus the interaction assumption is fulfilled. But the publicity and perception assumptions are also satisfied, for anyone who overgrazes is likely to be noticed and anyone who notices is bound to understand the collective and personal harm done. Thus we may expect, as the sanction assumption has it, that nearly everyone will disapprove of anyone else's overgrazing and approve of anyone else's not overgrazing. Since there is no great cost in not overgrazing, at least if enough others also refrain, the desire which the motivation assumption postulates ought to weigh sufficiently with people to elicit a general pattern of restraint. The upshot will be a norm of not overgrazing. Nearly everyone will conform to this regularity. Nearly everyone will approve of nearly anyone else's conforming and disapprove of nearly anyone else's deviating. And this pattern of approval and disapproval will help to explain why nearly everyone conforms.

I hope that this example will serve to show that it is possible to have norms whose emergence and persistence are derivable in an attitude-based way. The possibility is significant. With a norm like that of not overgrazing, a behavior-based derivation, or at least one which relies on tit-for-tat, is unlikely to be persuasive. The predicament is a type A dilemma, in which the lone defector will not put anyone below the baseline

50. Other things may not be equal with free rider problems such as those raised by truth-telling and promise-keeping, where the action type in question often represents, not just the cooperative option in such a many-party dilemma, but also the cooperative option in a two-party dilemma with one's interlocutor. See n. 33 above.

51. Garrett Hardin, "The Tragedy of the Commons," *Science* 162 (1968): 1243–48.

52. Taylor, *The Possibility of Cooperation,* pp. 26–27.

of universal defection. Why, therefore, should the potential defector or free rider expect others to stick with tit-for-tat? In particular, why should he expect them to risk all they have achieved and punish his lone defection by defecting themselves in response? The availability of an attitude-based derivation with a norm of this kind is therefore something of significance. The availability of the derivation means that political theorists may have a novel basis for identifying resilient norms, social theorists a novel hypothesis for explaining the rise and fall of norms that have appeared in history.

In identifying feasible norms, or in explaining why certain norms emerged and persisted, the idea suggested is that we should look to see how far the norms involve action types that engage our five assumptions. If the action type provides a collective and personal benefit, as required by the interaction assumption, does it satisfy the publicity and perception requirements? If it does, is the action type such as to attract approval, its omission disapproval, as in the sanction assumption? Is it, for example, sufficiently undemanding on the individual for people not to shrink from such disapproval? And if the action type satisfies all those conditions, is it the sort where the desire which the motivation assumption postulates—the desire to have approval rather than disapproval—is likely to outweigh conflicting motives? These questions represent a miniature research program for political and social theory.

Consider a norm such as that which most of us would hope to find operating in juries: the norm of taking seriously the question of whether the evidence establishes guilt beyond reasonable doubt. The behavior required by that norm is hardly independently motivated as a behavior-based derivation would have to suppose. And so the question of whether we can really rely on such a norm assumes some urgency. In dealing with a question of this kind, it will be useful to bear in mind the lesson of this article. We should explore the possibility that the norm is derivable in an attitude-based way. If it proves to be derivable, or derivable given certain additional constraints, this will reinforce our attachment to the institution of the jury, perhaps guiding us on particular issues of reform. If it proves not to be derivable in this way, then that raises doubts about the whole institution, at least for someone in the rational-choice tradition.

In fact, I would suggest, the jury norm does promise to be derivable in an attitude-based way. The activity required is collectively beneficial and in some measure beneficial to nearly everyone individually: it reinforces an institution from which almost everyone stands to gain. Thus our first assumption is fulfilled in this case. And so, more obviously, are the other four: everyone embracing or resisting the behavior is subject to the publicity and perception of other jurors; everyone is subject therefore to approval or disapproval; and everyone has a potential motive to display the behavior. Would the motive be effective, if actualized? We may hope so, especially given that juries are vetted to eliminate those with an interest in the outcome and that jurors are relatively protected from the threats of those with such an interest.

The vindication of the jury norm that I have sketched foreshadows other possibilities. It may be, for similar reasons, that norms of serving the public interest are feasible, or can be made to be feasible, in the realms of bureaucracy and academia. After all, the promotions committee, the lynchpin of such organizations, closely parallels the jury. And it may be too that professional norms, such as those to which doctors and lawyers subscribe, are or can be made resilient in a similar manner. Again the norms that have to be exemplified if patterns of self-regulation are to operate successfully in business and industry may prove to be feasible, in the light of our analysis, under appropriate conditions. Finally, and less congenially, the attitude-based strategy of derivation may enable us to understand why certain subgroup norms that are inimical to the large society—the norms of manipulative elites or criminal subcultures, for example—prove so enduring that no policy-making initiative should assume they can be displaced.

The possibilities are tantalizing. They make an interesting research agenda for political theorists who are concerned with which attractive norms are feasible, which unattractive norms inevitable, and under what conditions. And equally they point to an agenda for social theorists whose primary concern is explanation rather than evaluation. The fulfillment or nonfulfillment of assumptions like those listed may be very important in explaining the emergence or nonemergence, the persistence or nonpersistence, of the norms that interest social theorists. For the explanatory questions teem. How important a factor is size in affecting publicity? How far does publicity matter if the agent remains anonymous? Does popular understanding of the benefit or damage attending a certain activity—say, the damage done by smoking in public—encourage the appearance of a suitable norm? Does group conflict in a society—say, on a Left-Right or feminist-nonfeminist axis—undermine common norms such as those that we might expect to govern appointments and promotions? Does it mean that patterns of approval shift, for example, or that people come to care only for the approval of their own group? I hope that by adverting to issues like these I can at least signal the possibility that the attitude-based strategy for deriving norms is of more than just philosophical interest.

ONCE MORE, WITH COMMON BELIEF

It has been fashionable to argue that norms require not just the fulfillment of conditions like those given in the first section of this article but also the common belief that such conditions are fulfilled. People each believe that they hold, they each believe that they each believe this, and so on. Or at least they approximate to such common belief. Perhaps they have the belief but only *in sensu diviso.*[53] Perhaps they have the belief but only

53. Lewis, *Convention,* p. 66.

up to three levels.[54] Or perhaps they just lack at each higher level the contrary disbelief: they do not disbelieve that the basic matters hold, they do not disbelieve this, and so on.[55]

The reason for building a requirement of this kind into the definition of norms is that we would probably hesitate to describe a regularity as a norm if it were not fulfilled. Suppose for example that with a regularity, R—say, the regularity whereby people marry outside their families—the three requirements are fulfilled, but people do not generally believe they are: say, they generally believe that conformity is wholly explained by genetic predispositions. We might well hesitate to say in such a case that R was a norm. It would not be something that people thought it important for them to approve, since they would see their approval as epiphenomenal; thus it would not be something which served the ordinary role of a norm in their lives. Again suppose, a level up, that each person believed that the three requirements were fulfilled but believed that others did not believe this: say, each believed that others believed that conformity was genetically produced. Here too we would perhaps hesitate to say that R was a norm, for it would also fail to serve the ordinary role of a norm in the society: it would not be something each believed that others thought it important for them to approve.[56] Similar considerations would seem to carry weight, though progressively lighter weight, at higher levels. They suggest that norms do involve common belief, at least in the weakest sense that people do not disbelieve the relevant proposition at any higher level.

The requirement of common belief forces us then to tighten up our definition of norms.

> A regularity, R, in the behavior of members of a population, P, when they are agents in a recurrent situation, S, is a norm if and only if it is true that, *and it is a matter of common belief that,* in any instance of S among members of P,
>
> 1. nearly everyone conforms to R;
> 2. nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else's deviating; and
> 3. the fact that nearly everyone approves and disapproves on this pattern helps to ensure that nearly everyone conforms.

The question raised by this redefinition is whether the two strategies for deriving norms are capable of supporting a derivation of norms under this tighter analysis. I believe they can, and I would like briefly to show how.

54. Bach and Harnish, p. 269.

55. See Lewis, "Languages and Language," p. 166; and Gareth Evans and John McDowell, eds., *Truth and Meaning* (Oxford: Oxford University Press, 1976), pp. xx–xxi.

56. Notice that these considerations are not undermined by Tyler Burge's arguments in "On Knowledge and Convention," *Philosophical Review* 84 (1975): 249–55.

In arguing that conventions involve common knowledge David Lewis introduces the notion of a basis for common knowledge.[57] A basis of common knowledge that q is a proposition p such that everyone has reason to believe that p; p indicates to everyone that everyone has reason to believe that p; and p indicates to everyone that q. Given this, and given the mutual ascription of common information and inductive standards, p will indicate to everyone not only that everyone has reason to believe that p but also that everyone has reason to believe that q; and iterating again, not only that everyone has reason to believe that q but also that everyone has reason to believe that everyone has reason to believe that q; and so on. In such a situation it would seem reasonable to ascribe a common belief that q, at least under the negative construal of such belief. It is plausible that everyone believes that q, that no one disbelieves that everyone believes that q, that no one disbelieves that this disbelief is generally absent, and so on.

The notion of a basis of common knowledge suggests a nice way of showing that a derivation of a norm under our original definition also provides a derivation of the norm under the tighter analysis. This would be to show that the propositions involved in the derivation are the analogue of 'q,' providing a basis for common knowledge that p, where 'p' stands for the proposition that nearly everyone conforms to the regularity involved, nearly everyone approves and disapproves appropriately, and nearly everyone's conformity is ensured in part by that pattern of approval and disapproval. It turns out that this can be shown, or at least this can be made plausible, both for the behavior-based sort of derivation and for the attitude-based one.

Consider the propositions involved in the behavior-based tit-for-tat derivation of a norm of cooperating in some way with others. These are the crucial claims.

1. Universal tit-for-tat is a Pareto-optimal equilibrium.
2. Everyone adopts the tit-for-tat strategy, so far as it is the salient alternative.
3. And so everyone cooperates.
4. Universal tit-for-tat is also a coordination equilibrium.
5. Therefore everyone disapproves of anyone else's unilaterally defecting and approves of anyone else's defecting in this way.
6. And so, given that everyone is in a position to recognize the truth of 5, everyone has an extra motive not to defect unilaterally.

If we imagine a situation in which these propositions hold, then it is plausible to say that everyone there has reason to believe they hold; if he thinks about the matter, then he is likely to endorse the propositions or at least some less technical counterparts. Not only does everyone have reason to believe that the propositions hold, but the propositions also

57. See Lewis, *Convention,* p. 56.

indicate to everyone—they provide everyone with reason to believe—two distinct things: that everyone has reason to believe they hold, the evidence being equally available to all, and that tit-for-tat cooperation will satisfy the earlier conditions for being a norm, attracting general conformity and a general reinforcing pattern of approval for conformity. This means in turn, iterating, that they indicate to everyone, given that everyone has the same information and follows the same inductive standards, that everyone has reason to believe that cooperation will satisfy those conditions. And so on up the hierarchy.

That everyone has reason to believe that cooperation satisfies the conditions, that everyone has reason to believe that everyone has reason to believe that it does so, and so on, does not mean in itself that the common belief requirement is fulfilled. But it makes the requirement extremely likely to be met. It makes it likely, on our construal, that nearly everyone will believe that cooperation satisfies the conditions, that no one will disbelieve that nearly everyone believes this, that no one will disbelieve that this disbelief is generally absent, and so on. Thus we can see how a behavior-based derivation of a norm under the old definition can yield a derivation of the norm under the new.

The line of argument just run with the behavior-based strategy can be run also, as ought to be obvious, with the attitude-based approach. All that needs to be done is to replace the six propositions mentioned above with the claims involved in the four-stage derivation described in the last section. Thus we may conclude that the omission of the common belief requirement in our earlier discussions does not vitiate any of our results. What we did in defining and deriving no-frills norms, we can also do for norms in full dress.

CONCLUSION; AND A LAST OBJECTION RESOLVED

In conclusion, but before addressing one last objection, it will be useful to highlight the main claims made and defended so far.

1. Rational choice theory postulates that the things people generally do, whatever the basis on which they are chosen, are consistent with a major interest in economic gain and social acceptance; people do not generally flout such self-interest, even if they rarely think about it.

2. Norms are regularities such that nearly everyone conforms; nearly everyone approves of nearly anyone else's conforming and disapproves of his deviating; and this pattern of approval helps to ensure general conformity: whatever the basis on which people actually conform, the pattern of approval makes it unlikely that they will deviate.

3. Many norms will probably be inexplicable from a rational choice point of view. But it is still an important question whether we can identify certain norms such that, under suitable circumstances, rational choice theory predicts that they will emerge and/or persist. Such a rational choice derivation would identify those norms as significantly reliable; with some norms that will be good news, with others bad.

4. The definition of norms suggests that there ought to be two major styles of derivation available. One, the behavior-based strategy, would first explain the behavior of conformity and then explain the attitude of approval for that sort of behavior, given it is in place. The other, the attitude-based strategy, would first explain the attitude of approval for the kind of behavior at issue, whether or not it is in place, and would then explain the conformity in terms of a desire for approval.

5. The standard derivation attempted in rational choice circles is the behavior-based kind. The main examples are David Lewis's derivation of conventional norms and the derivation of tit-for-tat norms associated with a number of recent thinkers.[58] The Lewis derivation is impressive but the tit-for-tat variety has problems, at least for the many-party case.

6. Despite such problems, rational choice theorists have shied away from the other, attitude-based, strategy for deriving norms. They have been impressed, it seems, by the objection that people will find the giving of disapproval costly and will each abstain from the activity, seeking to free ride on the giving of disapproval by others to offending types of behavior. But while this objection may apply to the overt activity of disapproval—or indeed approval—it does not apply to the covert attitude of disapproval. The point is of relevance, because we care about the attitudes, including the unexpressed attitudes, of other people toward us, not just about their overt censure or punishment.

7. This observation shows the way to an attitude-based derivation of certain norms. There are five conditions such that where they are fulfilled rational choice theory predicts that certain norms will emerge and persist. The conditions seem to be fulfilled for some norms in every society—and for some norms in many social subgroups—and the derivation is therefore of practical significance. Just to mention two socially desirable examples, we are pointed toward a norm of not overgrazing the commons and a norm of conscientiousness in jury service.

8. There is a case for enriching our definition of norms, so that the conditions given are a matter of common knowledge. But the derivations discussed here can be extended to make such common knowledge also intelligible.

So much for the ground gained. To finish off our discussion, I turn to an objection: that neither sort of rational choice derivation can make sense of the fact that with many norms people are disposed to approve of conformity and disapprove of deviance on a moral or at least impartial basis, not just on a basis of self-interest. People overtly and covertly censure one another's failures on the basis that they are inimical to the common interest, are unfair, or whatever; they moralize about one another's transgressions. The objection is that rational choice theory cannot explain this and that it does not enable us to derive the presence of any moralized

58. These norms would count as conventions in Sugden's sense. They each represent one of a number of stable equilibria.

norms, as distinct from the norms that fit our comparatively undemanding definition.

The attitude-based derivation identified in this article enables us, happily, to counter the objection. Suppose that an unmoralized norm, N, is in place in a society or group, being explicable in a behavior-based or attitude-based way. It turns out that in that case there is an attitude-based derivation available for the norm of moralizing about N in the society or group: that is, for praising conformity and censuring deviance on an impartial basis, at least in a certain sort of context.

The five conditions under which we would expect a derivation to go through are fulfilled for any such moralizing. Everyone is better off if everyone else moralizes, since the N promoted by moralizing is to everyone's advantage, by the assumption that conformity with it is independently derivable; for similar reasons, everyone is better off in one respect for anyone else's moralizing in that way.[59] Thus the interaction assumption is fulfilled. The publicity and perception assumptions go through smoothly, at least for moralizing that is done in front of third parties. The sanction assumption goes through also, since each person will have reason to approve of anyone else's moralizing and disapprove of his failing to moralize, at least in an appropriate context. Finally, this should give each a motive to moralize that we may expect generally to be effective.

If there is a norm of moralizing about N in place then an offender need not expect everyone who notices him to offer moral censure: the context may not be of the sort appropriate. However, he is in a position to expect that the observer would censure him in the appropriate context, at least were costs low enough. And that means that he is in a position to know that the observer has an attitude of disapproval, specifically of moral disapproval, toward him for what he has done: he may not lay blame on him overtly, but he is bound to be covertly censorious.

This is a nice note to end on. It suggests that resort to the attitude-based strategy does more than extend the domain of rational-choice derivations, targeting norms that would otherwise be underivable. Resort to the attitude-based strategy may also deepen the reach of rational-choice derivations, enabling us to see why the norms derived may come to have a quasi-moral status in the relevant society or subgroup. This is not to derive the "ought" of morality from the "is" of rational self-interest. But it is to say that declaiming about what morally ought to be may be an activity which often makes rational self-interested sense.

59. On the point of moralizing as a practice, see Michael Smith, "Dispositional Theories of Value," *Proceedings of the Aristotelian Society,* supp. vol. 61 (1989).