

JOSEPH HEATH

THE STRUCTURE OF NORMATIVE CONTROL

One of the most commonly observed peculiarities of the instrumental conception of rationality is that when applied in contexts of social interaction it sometimes prescribes actions that will predictably result in suboptimal outcomes. Often these outcomes could be avoided if agents were able to credibly commit themselves to refraining from exercising certain options available to them. The prisoners' dilemma is the classic example. This problem has generated a small growth industry of attempts to modify the instrumental model in order to incorporate commitments. The reason that philosophers are so attracted to this project is that it seems to offer them an opportunity to finish off the job that Socrates began, viz. to refute moral skepticism of the 'rational egoist' variety.¹ None of this has worked very well, but enthusiasm for the project appears to continue unabated.²

But there is another, more important shortcoming of the instrumental model that has received considerably less attention from philosophers. This is the problem of indeterminacy. In a prisoners' dilemma, instrumental rationality seems to recommend that agents engage in a sort of collectively self-defeating behavior. However, in the vast majority of social interactions, instrumental rationality does not recommend anything at all. This is because all but the most

¹ This ambition is clearest in David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).

² Initial attempts to build commitment directly in the decision-theoretic foundations of game theory are now generally regarded as hopeless, so attention has turned toward eliminating *sequential rationality* as a constraint on equilibrium belief sets. This ambition is clearest in Edward F. McLennan, "Pragmatic Rationality and Rules," *Philosophy & Public Affairs* 26 (1997), pp. 210–258, but also underlies David Gauthier's strategy in "Assure and Threaten," *Ethics* 104 (1994), pp. 690–721.



simple games have multiple strategic equilibria, and instrumental rationality seems to lack the resources necessary to pick out just one. In many interaction problems, agents will therefore be incapable of “powering” their way through to a solution by systematically entertaining strategic hypotheses. This means that each agent will be unable to form rational expectations about what others will do, and so will be incapable of selecting a utility-maximizing course of action. Only in cases where the iterated elimination of strongly dominated strategies removes all but one of every player’s feasible actions is there a *procedure* that any player can follow in order to determine her best action.

The most famous game-theoretic solution concept – Nash equilibrium – gets around this limitation by effectively ignoring the question of how agents arrive at any particular set of expectations. What Nash saw was that a set of beliefs would *not* be acceptable if it contained a certain sort of defect, viz. if it was self-defeating.³ A self-defeating set of beliefs is one which, were players to actually adopt it, would lead them to act in a way that was contrary to the expectations contained in that very set of beliefs. What Nash then proposed, plausibly, is that it be a minimal condition on the credibility of any set of expectations that it not contain an intrinsic flaw of this type. What Nash left entirely open was the question of how agents are supposed to *get* to a belief set that satisfies this or any other constraint. And since Nash’s definition is negative, i.e., it specifies only what properties a solution should not have, it does not guarantee a unique solution. So naturally the question that becomes most pressing is how agents are supposed to coordinate on one or another of these sets. Nash thought that this portion of the theory would be fleshed out by future research, but there has been remarkably little progress on this front.⁴ In fact, Nash’s promissory note has been outstanding for so long that many theorists have simply forgotten that it needs to be redeemed.

Where there has been any progress, most of it has built upon Thomas Schelling’s early observation that certain psychological or

³ John Nash, “Noncooperative Games,” *Annals of Mathematics* 54 (1951), pp. 289–295.

⁴ See David M. Kreps, *Game Theory and Economic Modelling* (Oxford: Clarendon Press, 1990), p. 101.

cultural associations that agents may have with a particular outcome can make it “salient” or “focal” in a way that leads them to select it in a coordination problem.⁵ Schelling argued that such things as round numbers, equal shares, bright colors, etc., by attracting agents’ attention, could serve to favor outcomes exhibiting such properties. Similarly, an equilibrium might become focal by virtue of having been played before (an observation drawn upon by David Lewis as the basis for a general theory of convention⁶). Unfortunately, while Schelling provided a number of persuasive examples of this phenomenon, he did not attempt a detailed analysis. In particular, he did not suggest any account of how the psychological states induced by these “salient” properties were supposed to interact with the belief-desire states posited by standard decision theory. As a result, there has always been a cloud of mystery surrounding Schelling’s solution, because no one has been able to state clearly the intentional mechanism through which the results are achieved.⁷

In this paper, I would like to show how very simple deliberative microfoundations can be supplied for a Schelling-style equilibrium-selection mechanism, if one accepts the idea that psychological and culture factors can favor a particular outcome because they supply *non-instrumental reasons for action*. I will attempt to show that – contrary to the usual assumption – these non-instrumental reasons can be introduced without positing any dubious psychological entities or intentional states. However, I will also claim that these non-instrumental reasons need not only complement, but may also compete with, standard instrumental reasons for action. The advantage of understanding psychological and cultural factors in this way, I will argue, is that it enables us to account for the *force* of social institutions. Unlike the standard instrumental account, which must posit an essentially fortuitous coincidence of interests in order to

⁵ See Thomas Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960).

⁶ David Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969).

⁷ Robert Sugden has argued that the way agents label their strategies can generate focal effects, see “A Theory of Focal Points,” *The Economic Journal* 105 (1995), pp. 533–550, but he makes no attempt to explain how agents arrive at these labels. As a result, his analysis does not so much dispel the mystery as relocate it.

explain how social conventions arise, a theory that posits different *types* of reasons for action is able to explain how rules can exercise genuine constraint over the behavior of agents, and is therefore able to explain how social order can be achieved despite the persistence of underlying conflicts of interest.

I

The indeterminacy problem is disarmingly simple. In one form, the question reduces to one of how people manage simple tasks like passing on a sidewalk without colliding. There are two equilibria: I go left, you go right; and I go right, you go left. Furthermore, regardless of how either or us feel about left or right passing, we are both interested in avoiding collisions. This means that each of us wants to go left only if the other is going right, and wants to go right only if the other is going left. So how do either of us choose? Game theory, it appears, cannot tell us. This is the problem of indeterminacy. As Cristina Bicchieri puts it:

[Nash's] admittedly limited definition of mutually rational beliefs would be completely satisfactory were game theory just bound to define what an equilibrium is and the conditions which make it possible ... Yet normative game theory's aim is to prescribe actions that will bring about an equilibrium, which means providing a *unique* rational recommendation on how to play. Indeed, if the task of the theorist were limited to pointing to a set of rational actions, the players might never succeed in coordinating their actions, since different agents might follow different recommendations. Thus a unique rational action for every player must be recommended, together with a unique belief about the behavior of other players justifying it.⁸

For a variety of reasons, the majority of game theorists have given up on the idea that the set of equilibria in games can be significantly reduced on the basis of constraints that flow from the instrumental conception of rationality. Initially, it was thought the introduction of *refinements* on the Nash solution concept would significantly narrow the set of equilibrium outcomes. Solution concepts like Reinhard Selten's subgame-perfect equilibrium appeared to cut down the number of equilibria by focusing on the sequence

⁸ Cristina Bicchieri, "Strategic Behavior and Counterfactuals," *Synthese* 76 (1988), pp. 135–169 at 138.

in which moves were made.⁹ It soon became apparent, however, that these refinements did not significantly narrow the set of equilibria, they just controlled for the fact that the type of Bayesian reasoning employed in the Nash solution does not place any constraints on players' responses to zero-probability events. As a result, it is possible for Nash equilibrium belief sets to contain obviously false counterfactuals. All that the refinements do is eliminate these. As a result, they have no purchase on the usual type of equilibrium-selection problem.¹⁰

All of the other proposals that were intended to resolve the equilibrium-selection problem wound up exacerbating it. Some theorists maintained that equilibrium-selection was only a problem in "one shot" strategic games. Given the opportunity for repeated play of a particular game, it was felt that players would be able to zero in on a particular outcome and maintain that equilibrium in subsequent play. Further investigation showed, however, that no such expectation was in order. In fact, repeated play of the same game simply gives players the opportunity to develop more sophisticated strategies, ones that can prescribe a complicated pattern of different actions at different stages. Furthermore, in repeated games it becomes possible to sustain all kinds of action that would be out of equilibrium in any single instance of the game. The problem that this creates for equilibrium-selection is summed up in the so-called "folk theorem", which shows that the set of equilibria in every repeated game is infinitely large.¹¹

The last great hope was that the introduction of some kind of communication or signaling system would help players select

⁹ See summary in Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4 (1975), pp. 25–55.

¹⁰ Ultimately it created more problems, by embroiling game theorists in debates over the probability of counterfactuals. The so-called 'forward induction' problems, such as Eric van Damme, "Stable Equilibria and Forward Induction," *Journal of Economic Theory* 48 (1989), pp. 476–496, are just examples of cases where the probability of an equilibrium-sustaining counterfactual need not be equivalent to its conditional probability. For discussion, see Bicchieri, *supra*.

¹¹ Drew Fudenberg and Eric Maskin, "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* 54 (1986), pp. 533–554. Note that for such results to obtain players must not discount the future too heavily.

an equilibrium. The idea was that linguistic interaction could be modeled as a special type of multi-stage game, in which one player's choice of action allowed others to make accurate inferences about his beliefs or intentions. A communication system of this type could then be pegged onto a variety of standard games as a "pre-play" segment, allowing one or more players to announce their intentions before beginning the game. This idea was abandoned when it was discovered that models of this type also have the unfortunate consequence of increasing, rather than decreasing the number of equilibria.¹² Because the "meaning" of each player's signal is not fixed exogenously, but determined by the effective equilibrium, a new equilibrium can always be created by permuting the mapping of meanings to actions. Since the relationship between sounds and meanings is arbitrary, for every equilibrium in which "left" means left and "right" means right, there will be another in which "left" means right and "right" means left. As long as this relationship is common knowledge, no agent has any reason to prefer one over the other. This means that adding communication just expands the set of equilibria in any game.¹³

The view most widely shared by game theorists now is that some kind of additional component will need to be added to the instrumental model – some exogenous choice criterion – to produce an initial specification of which outcome should be expected.¹⁴ The textbook example of such a supplementary theory is Schelling's "psychological" solution. However, since Schelling's pioneering work, there has been very little progress made in determining the intentional mechanism through which this type of coordination is achieved. In order to correct this deficiency, I would like to provide

¹² Joseph Farrell, "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior* 5 (1993), pp. 514–531.

¹³ This assumes that meanings are not fixed exogenously. If they are, it undermines the Nash solution concept. For overview, see Joseph Heath, "Is Language a Game?," *Canadian Journal of Philosophy* 26 (1996), pp. 1–28.

¹⁴ Roger Myerson, for instance, suggests that the indeterminacy problem reveals 'an essential limit on the ability of mathematical game theory to predict people's behavior in real conflict situations and an important agenda for research in social psychology and cultural anthropology'. See Myerson, *Game Theory: Analysis of Conflict* (Cambridge, MA: Harvard University Press, 1991), pp. 113–114.

a reconstruction of Schelling's "focal point" theory that incorporates an explicit analysis of the belief-supports that favor the focal equilibrium.

What I would like to draw upon, in order to formalize Schelling's view, is a characteristic of the agent's motivation that usually gets ignored in decision theory. Decision theory distinguishes between three different classes of events relevant to choice: actions, states and outcomes. The action is an event that the agent controls, while the state is an event that is outside of her control. The outcome is then a third event that is produced through the causal interaction of the first two. In order to decide what to do, the agent needs to decide which of the actions available to her is best – we can say, to establish a preference ordering over the set of available actions. But according to the standard instrumental view, actions are not desired for their own sake, but only for the outcomes they can produce. This is to say, the agent does not have direct preferences over actions, only over outcomes. What instrumental reasoning amounts to is a procedure for projecting these preferences over outcomes back onto the set of actions. What allows us to do this is our beliefs. By telling us what the chances are of each state obtaining, our beliefs tell us what the chances are of any particular action bringing about any particular outcome. This allows us to develop a ranking of actions that reflects our ranking of the outcomes.

But what if the agent happens to like some actions more than others, independent of which outcomes they bring about? Actions and outcomes are just events – the former directly under the agent's control, the latter only indirectly so. If the agent prefers some outcomes over others, there is no reason that she could not also prefer some actions over others. Standard decision theory, however, tacitly eliminates this possibility through the set of axioms used to derive the agent's utility function. The von-Neumann-Morgenstern axioms, for instance, require that the agent be indifferent between any two actions that yield the same probability distribution over outcomes.¹⁵ Relaxing this restriction produces some interesting consequences.

In order to keep terminology clear, I will use the term "urge" to refer to an intentional state that generates a direct preference for

¹⁵ John von Neumann and Oskar Morgenstern, *The Theory of Games and Economic Behavior*, 3rd ed. (New York: John Wiley & Sons, 1953), p. 27.

an action, and “desire” to refer to one that results in a preference for an outcome. The question then is how to handle the interaction between urges and desires when constructing a utility function for the agent. We can assume that the agent starts with two basic preference orderings, one over actions, the other over outcomes:

urges: $a_1 \succ a_2 \succ a_3$

desires: $o_1 \succ o_2 \succ o_3$

Our goal is to merge these two rankings into a single preference ordering over actions. The first thing we need to do is bring in beliefs. These will allow us to transform the desire-based ranking of outcomes into a desire-based ranking of actions. Suppose we have the following beliefs (given as a set of causal relations):

beliefs: $a_1 \rightarrow o_3, a_2 \rightarrow o_1, a_3 \rightarrow o_2$

Thus allows us to generate a new ordering of actions:

desires: $a_2 \succ a_3 \succ a_1$

We now have two different orderings of actions, one derived from the agent’s urges, the other from her desires. In order to merge these, we will have to start by assigning cardinal values to actions within each list. This can be done simply by assuming (for the sake of argument) that desires and urges are experienced as having a certain intensity level. This will give us two vectors (payoff to a_1 first, a_2 second, etc.), say:

$U = (1, 0.5, 0)$

$D = (0, 1, 0.5)$

However, we cannot just add these two vectors together, because we have not yet provided any basis for comparison between desires and urges. One easy solution is to assign some kind of weight k to urges, that says how much they are worth relative to desires. We could then define the agent’s overall utility function as:

$V = kU + D$

If we assume that urges and desires are worth about the same in the above example, this will give us a composite ranking (which, it should be noted, is not the same as the ranking represented in *U* or *D*):

composite: $a_2 \succ a_1 \succ a_3$

In decision theory, whether or not the agent has a composite utility function of this type can safely be ignored. This is because there are no circumstances in which the two components of the function could come apart. In game theory, however, the situation is quite different. When strategic reasoning fails to generate a single equilibrium, it means that agents are unable to develop a set of beliefs about which state will obtain. This makes it impossible to say which actions will lead to which outcomes, and as a result, makes it impossible to project the agent's preferences over outcomes back onto the set of actions. For an agent with a standard utility function, this then leaves him without any reason to choose any action. But for an agent with a composite utility function, it just makes the desire component of the function drop out, leaving the original ranking in terms of urges intact.¹⁶

As an example, consider a case in which we hope to meet at a café, but where there are two places that we might go. If we both choose to go to the same one, we will be happy, but there is no way to tell in advance what the other will choose. This means that I know which outcome I prefer, but I do not know which action will bring it about. Absent the needed belief, instrumental reasoning gives me no guidance. So what should I do? One obvious suggestion is that I could simply give up on trying to coordinate and go to the café that I like best. In the original problem, my urge was subordinate to my overriding desire to meet you. But having no idea how to satisfy this

¹⁶ This can never happen in a decision problem, because the agent's beliefs are interpreted as subjective probabilities. This means that even when the agent has absolutely no evidence for or against the occurrence of any state, she can still fall back on Laplace's principle of insufficient reason – treat all events as equiprobable unless given reason not to. However, this recourse is not available in games, because using this principle would amount to ascribing a mixed strategy to the other player, which will either be false if it is not a part of an equilibrium strategy profile, or question-begging if it is.

desire, it simply drops out of the picture, leaving me with just the urge.

The important thing about these urges is that they provide a mechanism that can generate Schelling's psychological solution to the equilibrium-selection problem. Suppose my urge to go to café x is common knowledge. You can then guess that, when I give up on trying to coordinate, I will go there. This gives you a belief about where I will be, which then allows you to act on your desire to meet me there. But anticipating this gives me a belief about where you will go, and so the ability to act on my desire. In short, the urge gets converted into a stable pair of instrumental reasons, thereby selecting one of the two strategic equilibria. It is able to do this because it supplies a reason for action that is not caught up in the cycle of interdependent expectations that renders the strategic problem indeterminate.

The introduction of urges provides the basis for a very simple mechanism to generate Schelling's psychological solution. In multiple-equilibrium games, neither player begins with any instrumental reason to perform any action in the set of actions that belong to the various equilibrium strategy profiles. However, if player 2 has an urge to perform a particular action, we can say that she has an affective reason to perform that action. If player 1 knows this, then player 1 has grounds to expect that action. This gives player 2 grounds to expect player 1 to play his best response to that action, and therefore gives her an instrumental reason to play her corresponding best response. This in turn gives player 1 an instrumental reason to play his best response.

There is one point that should be noted about this solution. It is crucial that a sharp distinction be maintained between the two *types* of reasons for action. The solution only works if urges are not treated as simply an extension of the set of desires. This is why generating a composite utility function, as I did above, is actually a bad idea. Writing urges into the payoffs, in standard game-theoretic notation, just increases the value of some outcomes, without providing any basis for the development of rational expectations. In the café coordination game, for instance, using the composite utility function just increases the payoffs associated with one outcome, but does nothing to solve the coordination problem. It is precisely

because urges are *not* associated with outcomes, but with actions, that they are able to escape the regress of anticipations, and hence solve the equilibrium-selection problem. Writing urges into the payoffs gratuitously obscures the fact that some component of the agent's motive for performing the action is non-instrumental, and therefore eliminates an important difference between the two types of criteria that an agent could use to select an action. It is therefore important that we always keep two sets of books on the agent's motivations, keeping separate track of urges and desires. Furthermore, strategic interactions should be modeled using only the D "utility function" (above), not V .

Naturally, acting on the basis of an urge can always be re-described as some kind of instrumental reasoning. So one could say, for instance, that in going to my favorite café, I am not really selecting an action on its own merits, but rather selecting it for some further outcome, such as my enjoyment of the coffee, or the ambiance. But this is to miss the point entirely. The distinction between an urge and a desire flows from the distinction between an action and an outcome, which is in turn determined by the way that events in the decision problem are divided up into those are under the agent's control and those that are not. In the café example, my going to café x is an action, whereas my meeting you at café x is an outcome, since the latter depends upon factors outside my control, viz. your decision, in a way that the former does not. This distinction can be fairly *ad hoc* in decision problems (since nothing is ever entirely under my control, anything can be called an outcome). But in games the distinction is strict, determined through triangulation of the causal chain of events initiated by either player.

II

It is important to note that the psychological solution to the equilibrium-selection problem is still in many ways quite low-powered. To see this, consider the café-coordination game again. Without knowing where you intend to go, I have no instrumental reason for choosing one café over the other. However, if your favorite café is café x , then you would have an urge to select that café. If your urge for café x is common knowledge, this gives us both

instrumental reasons to go there. Unfortunately, if it is also common knowledge that my favorite café is café *y*, a problem develops. You might get smart, and instead of choosing your own favorite café, choose mine, and *vice versa*. The problem is then that you start out with an affective reason to choose café *x*, and I start with an affective reason to choose café *y*, from which you immediately acquire an instrumental reason to choose café *y*, and I acquire an instrumental reason to choose café *x*. These new reasons would then prompt each of us to switch reasons again, and so on *ad infinitum*. This means that urges are of only limited value in solving equilibrium-selection problems.¹⁷

If, however, both of us like the same café, then the problem disappears. (It is worth emphasizing that the problem does not disappear because the outcome of meeting there Pareto-dominates the outcome of meeting elsewhere, it disappears because our affective reasons for going to that café are complementary, and so generate a complementary set of instrumental reasons.) This means that the focal point solution only works on the condition that players' urges are complementary, only one player has urges, or only one player's urges are common knowledge. Since the latter two cases are somewhat marginal, Schelling focused his discussion on qualities of outcomes that are associated with very widely shared psychological propensities, under the assumption that they would provide complementary urges. (There is also evidence that people in coordination problems often do not act on their own urges, but on expectations derived from the ascription of "typical" urges to others.¹⁸) Many theorists, however, have found this to be too narrow a base to provide

¹⁷ Although the fact that the theory of action developed here takes urges to be exogenously determined means that it will need to be supplemented by some account of where this particular pattern of preference comes from. In principle, there is no reason why this supplementary theory could not include some account of how agents could rationally deliberate over and revise their urges. Thus the theory presented here leaves the cognitive status of the agent's urges an entirely open question (just as standard decision theory leaves open the question of where the agent's desire-based preferences come from).

¹⁸ This is how I interpret some of the results in Judith Mehta, Chris Starmer, Robert Sugden, "Focal Points in Pure Coordination Games: An Experimental Investigation," *Theory and Decision* 36 (1994), pp. 658–673. Other results, such as those which follow the 'rule of equality', suggest to me that agents are relying upon social norms rather than urges (see below).

an adequate account of coordination. They have argued instead that the need for a stock of complementary prescriptions to focus expectations provides an opportunity to explain the role of *culture* in social interaction.¹⁹

The “cultural” solution to the problem of equilibrium selection differs from the “psychological” solution only in that it takes the initial expectations to be drawn from a body of shared rules rather than a set of primitive dispositions. In the above example, a simple rule prescribing where to meet would eliminate players’ dependence upon the fortuitous coincidence of urges. This is because a rule system, unlike a set of urges, has the advantage of providing a set of *complementary* practical prescriptions. This is by virtue of the fact that a rule can either specify obligations for all parties, or else some combination of obligations and entitlements, that results in every agent having a favored action. It is therefore no accident that agents’ expectations converge – norms eliminate most of the guesswork involved in coordination through psychological propensities.

However, if norms are to provide a mechanism to generate equilibria, they must in some way provide agents with reasons for action. We can represent a social norm as a set prescribing an action for some or all players (or types) involved in an interaction. In order to be effective, the norm must be common knowledge among all players. An agent who decides to employ a particular norm in making a decision takes the portion of the norm that pertains to her and adopts it as a selection criterion. We can call such a criterion a “principle”, and the reason it provides as a normative reason. This gives us three types of selection criteria that the agent can use in choosing an action: beliefs, which take the hypothetical form “If you want o_x , do a_y ”; and both urges and principles, which take the categorical form “do a_x ”. It should be noted that the reasons provided by principles are similar to urges in that they cannot be represented as preferences over outcomes, since they attach directly to actions. However, the kind of resource that they provide in practical deliberation is quite different, since they provide essentially public resources for selecting actions. Thus it will be necessary, in order to implement this “cultural” solution, to keep three sets of books on the agent’s reasons for action.

¹⁹ See Ken Binmore, *Playing Fair* (Cambridge, MA: MIT Press, 1994), p. 140.

This bookkeeping exercise is important for reasons other than perspicacity. So far, I have focused exclusively upon situations in which affective or normative reasons for action can be *converted* into instrumental reasons. In such cases, the agent whose urge or principle provided the initial set of rational expectations was endowed with a set of redundant reasons for action. What has not yet been considered are cases in which these reasons conflict. What if dropping out the instrumental component of the agent's motives left her with an urge or principle that favored an action that was not part of an equilibrium strategy profile? The most typical case would be one in which an agent has promised to perform a particular action, but then finds that it is not in her interest to carry through. The "keep-your-promises" rule then provides her with a direct normative reason for selecting the action, while the desire for some other outcome provides her with an instrumental reason for selecting another.²⁰

Obviously a normative reason of this type is not going to generate a focal point. But more importantly, it raises the question of which type of reasons for action should win out. Either the principle or the desire is going to be frustrated, but which one? Presenting the psychological and cultural action theories as "equilibrium-selection mechanisms", and narrowing the relevant action sets in advance to those included in equilibrium strategy profiles, simply assumes that instrumental reasons should automatically trump other types of reasons (or that urges and principles should take effect only when strategic rationality fails). But there is no principled reason why this should be so.

What this analysis reveals is a kind of structural conflict within the will that cannot arise in ordinary instrumental reasoning. From a purely technical standpoint this is not a problem – it can be solved by just assigning some kind of multiplier to each consideration (as was done with *k* earlier) that assigns them each a certain weight relative to one another. The real problem that it raises is philosophical. We are committed to the view that the agent can be effectively motivated

²⁰ I am assuming that principles can directly motivate action, just like desires. In so doing, I am setting aside a certain traditional Humean line of argument. For the rationale, see Joseph Heath, "Foundationalism and Practical Reason," *Mind* 106 (1997), pp. 451–473.

by different types of reasons for action, and that these different reasons can, in principle, conflict. This leads us to the conclusion that an agent faced with incompatible instrumental and normative reasons for action is confronted with a *higher-order choice problem*, viz. how much weight to assign to the different sets of reasons that she could let determine her actual conduct.

The biggest mistake at this point, in my view, would be to suggest that the higher-order choice problem should be resolved instrumentally. One might argue, for instance, that since conformity to a “keep your promises” rule can give each agent access to higher payoffs than purely strategic reasoning (in prisoners’ dilemmas, for example), each has an instrumental reason to assign priority to the normative standard. Not only would this be faulty instrumental reasoning (since each agent does even better if he adopts an instrumental standard while the other adopts a normative one), but it simply begs the question. If the agent is trying to decide what standard of reasoning to adopt, he clearly cannot appeal to considerations that have force only relative to one of these standards. Instrumental reasons for reasoning instrumentally and normative reasons for reasoning normatively are question-begging, but from this perspective, so are instrumental reasons for reasoning normatively, and normative reasons for reasoning instrumentally.

In order to clarify this, it is helpful to draw a comparison between this higher-order choice and a similar choice involving the rate of time-discounting. The choice between satisfying an urge or satisfying a desire is structurally similar to the choice between satisfying a desire *now* or satisfying it *later*. In many situations, agents are faced with not just a single outcome, but a stream of future outcomes. In order to select a utility-maximizing course of action, the agent must have some way of comparing the present value that satisfying some desire at time t_1 has against the present value of satisfying it at time t_2 . The standard way of doing this is to introduce a discount rate, that reduces the present value of future payoffs at some uniform rate. This discount rate is normally assumed to reflect the risk that future outcomes will not be achieved, but also an element of pure time-preference. This is due to the psychological fact that, all uncertainty aside, future satisfaction appears to be worth less to us, at present, the further removed it is in time.

Informally, the way we discount the future reflects the way we balance our “short-term” against our “long-term” interests. Putting it this way might lead us to ask, naively, “What is the optimal rate of time-discounting?” or “What rate of time-discounting would a rational agent select?” But it is easy to see that this question is senseless if “optimal” and “rational” are understood instrumentally. There is no such thing as a utility-maximizing rate of time-discounting, because utility-maximization is defined as maximization of *time-discounted* expected utility. Thus every rate will be reflectively endorsed by the person who has that rate. This is trivially so, because it is in your short-term interest to favor your short-term interests, and in your long-term interest to favor your long-term interests.

So from the standpoint of an individual engaged in deliberation, it is senseless to ask what rate of time-discounting should be adopted. At best, agents will simply favor whatever rate it is that they have. Similarly, I would like to suggest that it is senseless for an individual to ask which type of reasons for action she should assign deliberative priority to. The question only makes sense relative to one of these choice standards, i.e., once this choice has been made. It is no surprise to then find that each choice standard will be reflectively endorsed by any agent who has adopted that standard. One will be better at achieving one’s instrumental goals if one assigns instrumental reasons for action deliberative priority, just as one will be better at fulfilling one’s normative obligations if one assigns normative reasons for action deliberative priority. If this is correct, then the agent’s fundamental choice disposition, like her rate of pure time-preference, should be taken as exogenously determined from the standpoint of practical rationality. Of course, to say that it is not determined by practical rationality is not to say that it is arbitrary. It just means that the agent does not choose it. This does not mean that it cannot be determined by environmental factors, such as evolutionary selection or socialization.

It is worth noting that from the strict first-person perspective, there is no reason for an agent to select one discount rate over another. However, if agents are in a position to promote particular discount rates *for each other*, there may be significant instrumental reasons for favoring a longer time-horizon. This is because agents who discount the future less heavily are able to sustain a greater

range of cooperative behavior in repeated games, which harbors potential benefits for all those who interact with them. Similarly, there are obvious evolutionary reasons why natural selection would favor organisms with a longer planning horizon. I do not intend to argue in favor of either explanation here, all I want to point out is that introducing a social and environmental perspective provides a number of resources for explaining the source of the discount rates that we employ.

Whatever its ultimate explanation, one thing that is perfectly clear is that socialization plays an important role in shaping our rate of time-preference. From a very early age, we are pressed to acquire discipline, to be less impatient, to defer gratification, and so on. This applies equally to our fundamental choice disposition. We learn to control our impulses, i.e., to subordinate our urges to our desires, but we also learn to respect the rules, i.e. to subordinate our desires to a set of shared norms.²¹ Successful socialization, I would argue, has the effect of producing individuals who weigh normative reasons much more heavily than instrumental ones, and instrumental reasons more heavily than affective ones. In the limit case, this takes the form of a lexical ordering of choice standards: normative, instrumental, affective.

III

There are a number of substantive accounts of the socialization process that could be used to supplement the formal analysis of rational action offered above. I would like to endorse one proposal that has occupied a very important position in theoretical sociology since Talcott Parsons. At the core of Parsons's analysis is the idea that socialization does not produce specific behavioral dispositions in an agent, but only a higher-order disposition to assign certain kinds of considerations deliberative priority over others.²² I have tried to show how the claim that socialization involves the acquisi-

²¹ See general discussion in Jean Piaget, *The Moral Judgment of the Child*, Trans. Marjorie Gabain (New York: The Free Press, 1965).

²² The classic statement of this view is in Talcott Parsons, *The Social System* (New York: Free Press, 1951). He refers to these higher-order choices as 'pattern-variable' choices, which select between different 'action-orientations'.

tion of a choice disposition of this type can be rendered plausible on the basis of formal features of the theory of action. For Parsons, however, the attraction of this formal conception of socialization is that it can be used to explain the somewhat mysterious role that *sanctions* play in sustaining organized patterns of social interaction.

The key idea in Parsons's account is the claim that what agents acquire through socialization is not just a propensity to assign normative reasons greater deliberative weight than instrumental, or instrumental reasons greater weight than affective, but also a disposition to punish those who fail to weigh deliberative considerations in the same way. This understanding of the agent's fundamental choice disposition provides a simple mechanism that can explain both how social order is maintained on a day-to-day basis, and how the dispositions needed to reproduce this order can be instilled in new generations of agents. In sociological terms, it explains the internal relation between socialization and social control.

It has often been noted that agents are, in general, punished for failing to conform to a variety of social rules. It has generally been thought that these sanctions are important in maintaining the orderly qualities of social interaction. The most obvious way of understanding these sanctions is to suppose that they directly maintain social order by providing agents with instrumental reasons for conforming to the rules. This idea underlies the Hobbesian "solution" to the problem of order. The claim, roughly, is that agents adopt a strategy profile that calls for sanctioning those who fail to conform to that profile. They are willing to adopt such strategies because the associated equilibrium is Pareto-superior to one that involves no sanctioning system. Thus the fact that agents conform to the rules, and the fact that agents punish those who fail to conform to the rules, can be explained from an entirely instrumental point of view.

However, there are a number of problems with this explanation that have never been satisfactorily resolved. Since punishing people is usually costly or inconvenient for the person doing the punishing, a free-rider problem emerges that undermines the integrity of the punishment mechanism.²³ Furthermore, reoptimization undermines the credibility of the threatened sanction – once the crime has been

²³ See Michael Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987).

committed, the deterrent force of the sanctions has failed, and so what is the point of incurring the costs associated with carrying out the punishment? Apart from these conceptual difficulties, there are a number of empirical problems with the attempt to understand social sanctions in purely instrumental terms. First of all, there is the obvious evidence that agents often conform to rules and sanction violations even when it is clearly against their interest to do so. Normally, it does not require a ‘policeman at the elbow’ to stop people from littering parks, stealing from each other’s vegetable gardens, etc. Furthermore, people often carry out punishments even when they have no real incentive to do so. Certainly there are times when people are being honest in saying ‘this is going to hurt me more than it hurts you’. In both cases there is a sort of commitment involved that cannot be explained in instrumental terms.

The more significant empirical observation, however, is that most social sanctions do not have any intrinsic punitive quality. Often all they do is signal a shift in attitude. For example, when an agent violates certain rules of the road and incurs the ‘natural’ punishment of a collision, the dissuasive power of the sanction is clear. But similar sorts of violations may lead to the “social” sanction of being honked, yelled, or gestured at. It is not obvious how these constitute punishments, since the sound of a horn does not in itself “harm” anyone. Similarly, when an agent says “I’m very disappointed in you” to a friend, this constitutes an extremely powerful social sanction, but is entirely symbolic in nature.

These kinds of anomalies are what inspired Emile Durkheim’s claim that sanctions do not so much enforce the rules as articulate and reaffirm them in the face of violation.²⁴ Durkheim attempted to explain social order by supposing that agents come to be motivated to conform to social norms by internalizing the sanctions associated with their violation. According to this view, after having been sanctioned a few times, agents will conform to a norm of their own volition, because failure to do so triggers feelings of guilt, shame or remorse. For similar reasons, when fully socialized agents fail to respect the normative order, the sanctions applied against them can be largely symbolic in nature, because they are used only to

²⁴ Emile Durkheim, *The Division of Labor in Society*, Trans. George Simpson (New York: The Free Press, 1933), p. 426.

activate these underlying feelings. According to this view, sanctions exist initially to punish, but punishment has a socializing effect by virtue of our capacity for internalization. This permits the punishment sequence to become abbreviated, and finally, rendered entirely symbolic, as internal control mechanisms slowly come to take over from external ones.

The problem with this view, however, is that it conflicts with certain observations about the kind of competencies that agents acquire through socialization. In particular, it appears that what agents acquire through socialization is not a commitment to specific patterns of behavior, but rather a set of extremely generalized dispositions. Agents seem to acquire a set of increasingly diffuse value-orientations, e.g. a desire to be a “good person”, to “get along with others”, and so on. This is evidenced by the fact that agents’ social competencies are quite flexible, e.g. they are able to substitute other behavioral patterns and adopt other cultural practices. Furthermore, the mechanism through which they make such substitutions involves a variety of explicitly cognitive resources (hence the difference between primary and secondary socialization).²⁵

Observations such as these led Parsons to suggest a conception of socialization according to which agents do not acquire a set of specific dispositions to perform particular actions, but rather an entirely formal disposition to assign a certain class of practical considerations deliberative priority. According to this view, agents are not punished for failing to conform to a particular normative pattern *per se*, but only insofar as their actions reflect a failure to assign normative reasons for action deliberative priority. Socialization designates the process through which *these* sanctions are internalized. Parsons used the word “deviance” as a technical term to refer to cases where agents adopt an inappropriate “action-orientation” in this way.²⁶ So according to this Parsonian conception, sanctions secure social order not by punishing simple violation of the rules, but by punishing *motivated* violation of the rules, i.e. deviance. This means that sanctions provide agents with an instrumental reason to follow the rules, but also, when internalized,

²⁵ See discussion in Parsons, *supra*, note 23, at 236–240.

²⁶ *Id.*, at 206.

they provide a disposition to assign normative reasons deliberative priority over instrumental.

According to this view, normative integration requires at least some agents who are disposed to assign normative reasons deliberative priority, and for those agents who are not so disposed, an effective system of sanctions to give them instrumental reasons for conforming to the normative pattern. The norm-conformative orientation is therefore understood to consist of both a disposition to assign normative reasons priority in one's practical deliberations, and *the disposition to punish those who do not*. Thus punishment of deviance is not a specific normative obligation, but a structural feature of the norm-conformative orientation. The second component of the view is that social sanctions, when implemented through a norm-conformative orientation, function simultaneously as a mechanism of *socialization* and *social control*. Social control refers to their instrumental significance – sanctions make deviance unprofitable, from an instrumental point of view. Socialization refers for their symbolic quality – through their expression of disapproval and the internalization mechanism, sanctions generate the disposition to assign normative reasons for action deliberative priority.

IV

This account of socialization has much to recommend it, but I do not want to insist upon that here. I would like to end with the more general observation that a multidimensional theory of rational action of the type that I have outlined, in which agents are thought to deploy three lexically ordered (or weighted) standards of choice, can be used to explain the possibility of both coordination and cooperation. An agent might fall back upon an urge to perform a particular action in cases where instrumental rationality fails, providing a Schelling-style solution to the interaction problem. But the agent might also, according to this view, cooperate in a prisoners' dilemma, because his respect for the norm that prescribes the cooperative action outweighs his desire for the outcome associated with defection.

The sort of considerations that might combine to recommend a particular action can be very complex. For instance, in a particular

choice problem, an agent could decide that a_1 is obligatory on normative grounds, then proceed to perform it. On the other hand, she might eliminate a_3 on normative grounds, but decide that both a_1 and a_2 are permissible. She could then decide that o_1 is the most desirable outcome, discover that a_1 will be the most efficacious in bringing about o_1 , and so decide on instrumental grounds to opt for a_1 . Both deliberative processes produce the same action, except that the former relies upon strictly normative reasons, while the latter combines normative and instrumental reasoning (and is thus “multidimensional”). Most social interaction will involve a complex combination of considerations, but insofar as one type of reasoning predominates, we can distinguish between three pure types of rational action.

To illustrate the distinction, consider again the case of two agents trying to avoid colliding when passing each other on a sidewalk. Where I live, there is a Schelling-type convention that pedestrians pass each other on the right. Its status as a convention is reflected in the fact that the pattern is not sanctioned. Most people have what I would call an urge to move to the right-hand side of a sidewalk when passing an oncoming pedestrian – it usually just feels more natural to do so. This minor cathexis I take to be the result of repetition. In traffic, however, there is a norm that prescribes driving on the right-hand side of the road. Even though the underlying strategic interaction already has the structure of a coordination problem, the dangers associated with a failure to coordinate means that the equilibrium is reinforced with an explicit rule. This is demonstrated by the fact that driving on the left-hand side is subject to very strong negative sanction (legally of course, but also by other drivers – one can safely pass an oncoming car on the left on almost any highway by using the shoulder opposite, but the other driver is likely to become extremely upset).

For an agent endowed with a fundamental disposition that provides an ordering of the three choice standards, acting in the context of system of social norms that provides the basis for a set of shared expectations, each social interaction problem can be given a determinate solution. However, in cases where such norms are lacking, the problem may remain indeterminate. This, in my view, is one of the strengths of this analysis – it does not prove *too much*. It

shows how institutions can come to be established and sustained, and thereby how coordinated and cooperative behavior can be secured. But it also shows how, in the absence of such institutions, uncoordinated and noncooperative behavior can persist. For example, theorists who have tried to show that it is straightforwardly rational to cooperate in prisoners' dilemmas seldom consider the fact that, if they are right, markets should not function (since price competition is an interfirm prisoners' dilemma). The theory outlined here is able to explain the persistence of certain suboptimal equilibria through reference to our decision *not* to normatively regulate certain kinds of interactions, e.g. market exchanges, and at the same time, to account for the fact that we are able to avoid suboptimal interaction patterns in other contexts, e.g. by forming orderly queues. Similarly, the fact that agents fail to coordinate can be explained through reference to the lack of shared culture and norms.

There is, of course, a substantial body of sociological theory that articulates precisely this understanding of the relationship between social institutions and patterns of instrumental interaction. However, most sociological theorists have been uncertain what kind of reasons for action social norms provide. This has given rise to a strong tendency to think of social norms as non-rational, which in turn creates a temptation to give them purely 'macro' level, or functionalist explanations. The analysis developed here provides simple deliberative microfoundations for the classic multidimensional theory, which will hopefully help bridge the gap between the theory of social action and the theory of social order. At the same time, it extends the basic conceptual apparatus of preference-based decision theory in such a way as to eliminate two of the most problematic results of game-theoretic analysis, without denying any obvious facts about the structure that strategic interaction often assumes.

*Department of Philosophy
University of Toronto
215 Huron Street
Toronto, ON, Canada
M5S 1A1*