

The Relevance of Decision Theory to Ethical Theory

Jan Narveson

Accepted: 21 July 2009 / Published online: 29 April 2010
© Springer Science + Business Media B.V. 2009

Abstract Morality for the purposes of this paper consists of sets of rules or principles intended for the general regulation of conduct for all. Intuitionist accounts of morality are rejected as making reasoned analysis of morals impossible. In many interactions, there is partial conflict and partial cooperation. From the general social point of view, the rational thing to propose is that we steer clear of conflict and promote cooperation. This is what it is rational to propose to reinforce, and to assist in reinforcing in society; it is not necessarily what it is individually rational to do. Even so, given the general situation, the rationality of its reinforcement will typically support the rationality of individual action as well. Game theory makes it possible to clarify these interactions, and these proposals for social solutions.

Keywords Morals/morality · Rationality · Rational action · Intuitions · Interaction · Game · Game theory · Morals · Morality · Cooperation · Defection · Reinforcement · Prisoner's dilemma · Chicken · Battle-of-the sexes · Pure coordination · Zero-sum

1 Introduction

Is morals rational? Is it even an open question whether it is? Evidently many think that the question is indeed at least “open”; but most classic writers have thought it obvious that it is not open at all—that rationality, the voice of reason, is also the voice of morality. We need to clarify that issue, surely. It is perhaps not too much to say that we all would *like* morals to be somehow rational. But here a crucial distinction needs to be noted. (a) It is generally supposed that we can be rational *about* morals without insisting that the underlying theory of morals is or is derivable from more general ideas of rationality, or still more specifically from what has come to be called rational decision theory—just as we can be rational about trees without supposing trees or tree-behavior to be rational. (b) Alternatively, there is a hallowed tradition in philosophy for the claim that morals is rational in the sense that its

J. Narveson (✉)
University of Waterloo, Waterloo, ON, Canada
e-mail: jnarveso@uwaterloo.ca

prescriptions are actually the prescriptions, somehow, of reason itself. John Locke, for example, says that “The *State of Nature* has a Law of Nature to govern it, which obliges every one: and Reason, which is that Law, teaches all Mankind...” (Locke 1690, para. 6). But the claim that reason “is that law” is simply tossed off by Locke, as if it were obvious, no further explanation provided.

The aim of this paper is to put the case for the second point of view—but improving, it is hoped, on Locke and the many others who talk as he does in the process. As one of those who, over the years, have been persuaded of the importance of rational decision/game theory for moral theory, it would seem to be a matter of professional conscience to try to get this straight. I shall suggest that game theory is a useful tool for making us appreciate the problem, and also for pointing out the direction toward such solution as there is—which I argue to be considerable.

We are rational about a subject if we get the evidence right and draw the best conclusions from that evidence, or supply hypotheses with maximum explanatory power on behalf of that set of evidence. Suppose that ethics is a matter of “intuition” or basic instincts, or alternatively of open-ended social influence, or even of unlimited individual choice. If the evidence supports one of those views of the matter, then we are rational to embrace that view. But none of them would make ethics *itself* rational in the sense in which so many theorists of morals suppose it to be. Thus in the first of these options, the response of an agent to some relevant situation would be untutored and definite: the agent would not have “reason” to do as he does, exactly, though there would, presumably, be a possible account of the causes of his reaction. But that would be all. One would feel let down if such an account were all that is available for morality, since those accounts would not provide a rational basis for the normative stances themselves. They would enable us to understand morality, but not to see it as rational in the relevant and proprietary sense in which classic philosophers and many contemporaries have supposed that it is.

Of course, the philosopher who resorts to an intuitional account of morals could claim some measure of creative potential by noting underlying inconsistencies or points at which our ordinary “intuitions”, as they are called, seem to collide either with each other or with known facts. Nevertheless, intuitionist type theories leave the fundamental rationale of morals impossible, unknowable—because, basically, nonexistent. The question of why the basic principles (or basic particular judgments, in some versions) should be thought to be true (or rationally acceptable in whatever alternative to the concept of truth the theorist may deem appropriate)—is simply not an askable question, on these views of the matter. At bottom, those judgments just *are* whatever they are; the intuitionist says of such judgments, “and that is *that*.”

No doubt he will add that we must, after all, start somewhere. True, of course: but do we need to start with *morals*? Can moral truths be derived only from other moral truths? They would if, as classic intuitionists maintained, rightness were a “unique nonnatural property.” But if morality is analyzable, then this claim is rejected. We would show that something is right by showing that it meets the requirements that the correct analysis of it lays down.

And indeed, the first thing to be said on behalf of morals is that that *isn't* “that.” When someone makes a moral judgment, in our view, she may be called to account for it. Why? Because morals is *interpersonal*. That a person may be unable to account for some value judgment of her own is possibly interesting, but also possibly not. Who, after all, cares? This feature of intuitionism, as I have just stated it, seems rather academic. It does not bring out what's really wrong with intuitionism in *moral* theory, the subject of this discussion. Morals is different. For when morals is at stake, it isn't just one's own behavior that is in question. There are other people out there, and morals purports to concern them as well as

oneself. Morals aim to provide rules for all, not for just one or some select few agents (or theorists!)

One recent writer tells us that “There are many good reasons not to take the mere fact of disagreement to tell against either the objectivity of ethics or the truth of any particular objectivist theory of ethics.” (Jeske 2008, p. 107). But the “mere fact of disagreement” is not what is in question. We do of course have disagreements about all sorts of things, many of them perfectly “objective” in any reasonable sense of that term. But intuitionism does not lean on any mere fact of disagreement: it instead makes disagreement in principle impossible to resolve by rational methods, if it comes to that. Maybe it never will, of course. The same author allows that “We have to wait until the end of the day, as it were, before we start worrying about persistent disagreement, and that ‘day’ may last years ... Even at the end of that day, we have no way of being sure that all have grasped the relevant concepts.” (Jeske, p. 75) But it isn’t a matter of time. Objectivity does not mean that the issue can be resolved in 10 min. And after all, there are plenty of issues where we have, as we might say, all the time in the world. I am quite sure that there are many people whom I will never be able to persuade of the virtues of Haydn’s string quartets, if persuasion were needed. But a disagreement about whether killing innocent people is wrong is of a rather different order. We—society—can live with the first sort of “disagreement,” no problem. Not so the second. But if the wrongness of such killing is in principle a matter of a nonderivative “fact,” not even amenable, in principle, to analysis and the bearing of this or that objectively ascertainable fact on the matter, then something has gone very wrong. We can *not*, I think “live with” *that*.

Because our behavior towards each other affects persons with minds of their own, beyond the agent, and because putting our behavior into a moral framework means considering it in public, opening it to public scrutiny and reaction, there is a problem with intuitionism that goes beyond the merely academic. Suppose I propose to treat you pretty badly in some respect or other, and you want to know why I’m doing this—and all I’ve got to say is, “Gee, I don’t know—I just seem to have this voice saying “shoot this person!” You will understandably be not very happy, and of course you may reach for your gun, if you have one, or take steps to acquire one if you do not. Morals aims to improve human relations by requiring that there be reasons for our judgments, reasons necessarily accessible to, discussable by, and in the end acceptable to all. Hopefully, erroneous judgments can be shown to be such, the right ones confirmed, or at least made plausible, by a genuine analysis. This is where the rational-decision approach to morals seems to offer promise. We can also add that this sort of claim looks very much to be the assumption embedded in the very idea of morals. Morals is thought to be common sense, common reason, evident to all. The theorist might take (and many have taken) the view that this assumption is actually a delusion, and that this “very idea of morals” is simply not, on closer inspection, one that can be fulfilled. On too many familiar views of morals advanced by philosophers, especially in the last century or so, such a theorist would be justified. But modern game-theoretically oriented moralists hold that those familiar views—such as intuitionism, emotivism, “particularism,” and, despite its misleading name, moral realism—should be rejected as basic accounts of the subject. They bark up the wrong trees. Much more promising is the proposal that what we call “morals” are *agreed* principles for the regulation of our mutual behavior. If we agree, that helps: each party can attempt to make his case by appealing to principles common to all, because undergirded by reasons that all can understand and do actually accept. That’s the hope, anyway.

Defenders of the rejected theories may claim that they are being caricatured or misunderstood in the above. They may claim that somehow we all have the *same*

“intuitions” about what we all ought to do. But the very idea of intuition creates a problem at this point. For we may *say* that all have the same intuitions, and yet it often appears that we don’t. It “appears” in that some people say one thing, others the opposite. Now, the intuitionist at this point wants to insist that one of them is wrong. But what resources does he have available for showing this?—since on his own view, we *cannot* “demonstrate” in the sense of appealing to further reasons or evidence from which the correctness of the one view and the errors of the other can be shown. That is, after all, the very idea of intuition—that it is “intuitive” *rather than* a matter of deduction from more fundamental premises. On the intuitionist’s official view, at some point there *are* no more fundamental premises.

Recent writers on morals tend to use the term ‘intuition’ with less epistemic baggage. An intuition is what we judge, without knowing further why, though also without assuming that there is no further “why” about it. Intuition in this sense is simply judgment. But it has no privileged epistemic status. It does not in any way deny the possibility that further work might be done to show *why* we might tend spontaneously to judge a certain action right or wrong. ‘Intuitionism’ in this sense simply is not a moral theory, and should be left to one side for these purposes.

Of course, there is discussion and discussion. There is often said to be no disputing about tastes, and it is almost as often noted that despite this, people do in fact dispute, often at great length, about such things. But the point of the maxim is that we mustn’t expect the discussion to lead to agreement, or perhaps to much of anything other than the passing of the time of day. And when it’s a matter of aesthetics, that is plausible. But things are different when it comes to *moral* discussion. Hume famously says, “The notion of morals implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it” (Hume, *Inquiry*, IX.1). The way Hume puts it raises a question: if indeed morals is a matter of “sentiment,” then how are we to be so sure that it will indeed be “common to mankind”? We might expect that if the “agreement” in question is really to be *expected*, morals should be a matter of something more like objective fact. However, that morals implies some *basis of reaction* that is “common to all mankind” is indeed plausible. If we look further, perhaps we can see why *we*, even though not you or I taken only by ourselves, should come up with a pronouncement in favor of or against some kind of action.

Hume appealed to a moral sentiment, and that raises the problem that sentiments seem to be variable, as well as not obviously rational. I said that we seem to need something more like “objective facts”; but the trouble with that, in turn, is that objective facts look to be the wrong sort of thing for morals. Again it was Hume himself who famously insisted that moral beliefs need to be able to *move* people, and yet objective facts, just like that, look to be fundamentally inert. If facts are to move us, it will have to be because they are appropriately conjoined with something like sentiment, after all—with emotion, desire, impulsion. All such things seem to go beyond facts. But in doing so, they look also to present the threat of a sort of subjectivity which seems, again, to be the undoing of rationality. Where, then, do we go from here?

2 Morals

There is a decent answer to that—fortunately—and game theory turns out to be a, or perhaps the, key tool in formulating it. But to see this, we must first emphasize the distinction I have been implicitly leaning on, between two general sorts of evaluation. On

the one hand, we all make decisions of all kinds, and we make them in light of our values, whatever they are, or if you distinguish these, in light of our values, desires, or interests. But sometimes we decide in a social context, and specifically in matters regarding which our actions are likely to affect others in various ways. Some of those ways will make little difference, but others will make a fairly enormous difference to those affected, and they in turn will be moved in consequence to respond in ways that affect the original agent. These responses of others put us in a new situation. They make a difference, varying from trivial to fatal. Further, they introduce a major dimension of human experience: communication. In being able to anticipate the behavior of others, and they ours, there is the possibility that we will be able to communicate in ways that enable us to do better. We enlist the assistance of others, and they ours, in a cooperative manner—or, unfortunately, sometimes quite the reverse. Game theory is a major resource for this reason. Different agents coming into relation with each other may be able to analyze their interactions in a productive way. Such analyses may also suggest the possibility of developing general guides to conduct—principles, perhaps. That game-theoretic analysis helps us to develop insight into such ideas as moral principle, virtue, self-restraint, and the like, may come as a bit of a surprise. But that is the claim advanced here. It is *not*, however, the claim that moral theory simply *is* decision-theory. I do not propose that the rationally best action, at least insofar as it is possible (when it is) to track that with game-theory tools, is ipso facto the morally best action. I deny that, and think that game theory enables us to see why we would want to reject it.

Morals isn't, then, just any old set of evaluations. Moral judgments and principles and rules are aimed at adjusting behavior toward *other persons*. Historically, to be sure, there is also a wider use of the term in which its subject matter is the much more general subject of *how to live*. But though that subject is of obvious importance, it is much too general to identify the subject of morals. That subject is crucially concerned with the fact that there are other people in the world, people we deal with, affect, and are affected by. Those people are independent decision makers. Notoriously, people differ, and in particular they differ on the subject of how to live. Those differences are both extensive and deep. Yet we encounter each other, and the question arises how we are to cope with the various differences such encounters are bound to bring to the fore. That is the project of morals: to serve as guides or rules for general interaction among miscellaneous people—not just ones that share our own tastes or lifestyles, or who are close genetic relatives or fellow members of some club. When a moralist proposes that some sort of action is right or wrong, he supposes himself to be speaking to just anybody (—almost?—See below), rather than some narrow, special subset of the human species. Given the differences among persons, is this possible? That is to say: is it possible to find good reasons why we should, in general, approve of and encourage some kinds of actions on the part of just anybody, and disapprove and discourage others? Or are we compelled to remain in the swamps of sheer sentiment or intuition? Game theoretic analysis offers a means of making sense of these kinds of judgments that raises them above that level. When we claim that somebody acts wrongly, we are indeed, standardly, expressing disapproval of what he does—but, we hold, there is a basis for this that is lacking in the case of judging the taste of chocolate.

Our topic is “The Relevance of Rational Decision and Game Theory for Ethical Theory.” But in this discussion, I restrict the topic to *morality*, taken in what might be called a “social sense.” Decision theory is relevant (insofar as it is) to the making of decisions about virtually anything, but morality is a special case. I take it that morality, in this special—but wholly familiar—sense is a bundle of directives, addressed to people generally, purporting to lay down requirements or at least strong recommendations for the behavior or anyone

and everyone. But certainly there are much broader senses in which the term has been used. Aristotle, for example, identifies the moral with the department of “passions and actions,” claiming famously that the right way is to aim at “the mean” in respect of them; only in Book V of his *Nicomachean Ethics* (Aristotle 1941, 952–964) do we get to the specifically interpersonal, officially. Why, then, confine the present inquiry to the former, apparently much narrower, sense?

Note first that it is not narrower in the sense of restricting the *content* of the possible sets of requirements so called. For it is certainly possible that a society will have a general rule requiring people to “aim at the mean,” for instance—or a rule prohibiting it, for that matter. There have been moralities that attempt to control people’s religions, their eating habits, and so on indefinitely. So the way in which the notion of morality intended here is “narrower” is that it intends universality in a way in which individuals’ decisions about what to do need not aim at any such thing. I can be quite conscious that others would make entirely different choices than I would in a given choice-situation, without *necessarily* thereby thinking that the others are doing anything “wrong.” I may not care what they do in those respects—right and wrong needn’t come into it at all. The right choice of wines, of moves in chess, of operas to attend, of persons to befriend, and so on, *need* not be thought of as moral at all, in that first sense. All of them, of course, do come under a broader category. For this latter purpose I like the term, “theory of life”, or *how to live*, intending thereby not just grand philosophical pictures about that, but everything in the way of what might be chosen or not as one lives one’s life. As soon as someone takes his or her ideas about this and makes them into canons for others, that person would be making those values into *moral* values. But he or she need not do so.

There is reason to inquire into morality in this narrower sense: for morals as a social “institution” (or perhaps we should say, a sizable set of specifically differing such institutions, varying considerably from one “society” to another) has or can have a powerful impact on any given individual’s life. All of us come in touch with others, frequently, all but continually, and how they respond to our behavior matters. A society’s legal and political institutions, of course, obviously have a huge impact on our lives, but so too do its various moral codes. Moreover, it is plausible to hold that political and legal institutions are themselves subordinate to moral considerations. We can believe that some political initiative or some feature of the current law is susceptible to severe criticisms of a moral kind—that it violates some people’s rights, for example. How much effect this might have is variable, but certainly it can be enormous.

Of course this distinction is conceptual: we can understand that there is a difference between something’s being a social rule and its being only an individual choice or even an individual rule of action, and thus can raise the question whether this or that *should* be a social rule even if it is not currently, or should not be even if it currently is. And we can do this without any necessary implication that individual values in general belong in the former.

It is quite another issue whether morality should be understood to derive from certain general *values*, taken to have a status such that to incorporate *those* into the moral code of society at large is *ipso facto* the thing to do. Game theory has no use for this idea, in general. It analyzes the behavior of possible, rational individuals, who espouse the values in question—or don’t, as the case may be. For game theory, in other words, values are wholly individual. If there is a basis for a social rule, then, it will have to be found in some kind of commonality among individuals. Just laying it down that something or other is “good” will do nothing for individuals who deny that it is so, and who act accordingly.

We can here make use of a distinction somewhat obscurely made by Jean-Jacques Rousseau, between what he calls the “General Will” and the “Will of All.” The latter,

Rousseau says, is merely the “sum” of all the particular wills in society. This we should expect to be a jumble, since for many or perhaps even all randomly chosen pairs of persons, A and B, A is likely to want something such that, if he gets it, then B will be rendered unable to get something that B wants: which is to say, their wants are mutually incompatible—in some degree in conflict. If there is to be anything of the sort that Rousseau means by the “General Will,” then, we will need, as he says, to somehow iron out the pluses and minuses, and come up with a Common will, which would aim at something that could in some significant way be said to be directed toward the *common good*. (Rousseau 1978, p. 61.) But how is that to work? Conceptually, it seems possible to interpret this in only one way: that everyone’s wants are to be *consistent with* everyone else’s. And everyone is to embrace the General Will, totally, so that we can say that A wants what he does if and only if it is compatible with a general rule to the effect that you are to want something only if it is possible to get it without thereby frustrating the will of B, who in turn ... etc. In short, everyone’s wants are connected with everyone else’s by a chain link, as it were, so that each person wills only what is mutually compatible; the elaboration of this would provide the content of the Social Contract. The bottom-line criterion is mutual compatibility.

So stated, Rousseau’s idea is not easily realized. This is for two reasons. First, of course, there is the difficulty of formulating his idea in a way clear enough so that we can be sure it actually does imply anything at all in the way of general rules of behavior clear enough to be followed (or violated) by anyone. Second, and perhaps worse: Rousseau requires that each person, faced with a choice between pursuing one of his desires that turns out to be incompatible with others, and therefore does not meet his criterion, and another choice that is so compatible, will thereby and always prefer the latter to the former. The trouble is, it surely seems as though the former desire could be *stronger* than the latter. Rousseau can hardly think that everyone’s behavior, even in a civilized society, automatically does meet his requirement. So where do we go from here?

Game theory, as I will illustrate, shows us how this problem can be clarified. It is not clear whether it shows us how it can be solved. But the idea aimed at by moral theorists of this general persuasion is that game theory paves the way to understanding how—that is, in what way—morality is rational.

This has not escaped the attentions of other classic philosophers. Hobbes, Locke, and Kant, for example, were champions, in principle, of general liberty. Kant, for instance, tells us that the *Universal Principle of Right* is this:

“Every action is right which in itself, or in the maxim on which it proceeds, is such that it can coexist along with the freedom of the will of everyone in accordance with a universal law.” (Kant 1965, p. 35).

The words ... “*can coexist along with the freedom*” of others make it clear that compatibility is (being claimed to be) the bottom line.

How do we get from here to there? ‘Here’ is the rational calculations of individuals seeking what they take to be maximum benefit, while ‘there’ is the above principle; calling upon us all to refrain from pursuing our advantage *at the expense of others*, that is, in such a way that our pursuit of it essentially undermines someone else’s pursuit of it. It is hardly obvious that those two general conditions are even compatible, let alone that the first somehow leads to the second. On the other hand, those who think it obvious that they are not compatible speak much too quickly.

Here we need two things. One is, roughly speaking, a theory (sketchy, inevitably) of human action, and the other is, roughly speaking, game theory.

The “theory of human nature” in question is simple, in the abstract terms needed for discussion at this level. We are concerned with humans and their interrelations. So what are these entities? That is, what general features of them matter here? Three properties are relevant:

- a) a set of *interests* (or *utilities*, or *values*)
- b) a capability of gathering information and drawing inferences: Humean rationality, in effect
- c) an ability to communicate sufficiently well so as to make complex, and especially strategic, interaction possible. That is, A must be able to adjust his behavior on the basis of his perceptions of B’s behavior: what A does will be some function, in part, of what A thinks B will do.

We may class the development of game theory either as a part of (b) or as a separate thing. Notionally, it is a part of (b): game theory is intended to be inferential, mathematical. As such, it aims to draw out the implications of various packages of abstractly characterized interests and circumstances, and in course to yield, perhaps, plausible directions for the inquiring agent who wants to know what to do.

2.1 Morality—Not Assumed

Not on this list is morality, nor anything specifically equivalent to it. People typically do have moral senses, in their adult years at least, and a hallowed tradition has it that this sense is “innate.” But to posit innateness for it is problematic, since obviously many people do not behave as if their actions were controlled by moral considerations. If it is claimed that there is “moral law,” it at least cannot have the status of the laws of nature, in the sense of that expression in which the law of gravity is an example, say; for—obviously—people fairly often do what is wrong. Nor is it evident that *everyone*, even among adults, is equipped with a moral sense at all. A safer assumption is that any moral senses we may end up with have come about by socialization, and perhaps also by internal reflection, possibly even rational reflection. Philosophers thinking to confirm or endorse some substantive theory of morals must think that the latter at least *can* play a strong or even decisive role, in the end. By reflection, perhaps, we can confirm the various conditionings that socialization consists in—or perhaps disconfirm and undo them. The latter possibility can hardly be ruled out in advance, and the history of philosophy has certainly had its share of cynics, doubters, rebels, and revolutionaries. It can hardly just be assumed that such persons are less than rational.

On the other hand, it is assumed that we are able to influence each other in various ways. People can threaten, cajole, show by example, and so on. These possibilities run to some degree independently from the other interests of agents. We can react not only to particular actions of others, but also to their perceived dispositions and tendencies. Morals involves utilizing such behavioral reinforcements. In that sense it is like government, which exerts control via threats backing up rules, where those threats have no necessary connection to the other interests of those governed. But a rational morality will utilize those threats in ways that make sense given the ways in which we interact.

2.2 Rationality

We next need to say something about rational action. If morality is to be rational, we need to have a decent idea of what rationality is, in this domain. But it needs to be rationality in a recognizably ordinary sense of the term, rather than some special sort of rationality, question-beggingly identified with the special theory of morals that we want to propound.

What, then, is this general kind of practical rationality? A standard characterization is supplied by Rawls, who says that “the concept of rationality must be interpreted as far as possible in the narrow sense, standard in economic theory, of taking the most effective means to given ends ...” (Rawls 1971, p. 14) whatever those ends may be. An agent is rational, then, if she chooses his actions in such a way as to best promote *what she values most*. Is that a matter of “effective means to given ends”? Well, for one thing, the ends are *not* “given” in the sense of fixed and beyond reconsideration. An agent at any given time has an end “in view,” but many things could, and often do, make him change his mind. Nor will he necessarily see his actions as a “means” to some externally definable end. To take an obvious example, the dancer wants to move in certain ways: her actions aren’t obviously “means” to some *further* end than that of moving in just *those* ways. For her, the “means” would be such things as the rigorous training previously undergone in order to enable her to do those things. Deciding what to do is usually very largely a matter of deciding which actions will conduce to independently definable ends, but it is also to some extent not so; and at times, hardly at all so. The notion of rational behavior, then, should not be understood “narrowly” in that sense. But we can talk of the individual’s preferences, whether or not they in turn easily break down into talk of means and ends. And, we can always understand someone else’s behavior—insofar as we do understand it—as issuing from his preferences—his interests or values, insofar as we can find out what those are.

Are values the same as wants? We appear sometimes to value something, yet say that we do not want it. And sometimes, we find ourselves wanting something, yet we claim, to ourselves, that what we want is bad in some way, and we shouldn’t have it. All this is true—but it doesn’t necessarily *matter* for moral purposes. For those purposes, our question is where people stand, what we can expect them to *do*. When they make evaluative statements, our question is how these impact on the behavior we can expect to be forthcoming from them. *One* way in which we can have a disconnect between wants and values is when the values in question are moral ones. We may very much want to do x, but upon reflection, we might see that x is something that we can’t expect to be allowed by our fellows—nor, in the end, by ourselves, taking our situations fully into account. Wanting what is *morally* wrong and not wanting to do what is morally right are perfectly conceivable, and indeed fairly normal.

But the idea of morals is to correct, override, our personal values. A moral rule tells us to do things we might not (otherwise) want to do, and to refrain from things we might (otherwise) want to do. If all of our values were compatible from the start, morality would be unnecessary and completely useless. The constraining, or overruling, function of morals is what makes it interesting from the theoretical point of view. For after all, how *can* we “override” our interests, since the latter are all we’ve got?

Note that I use here the normative words ‘best’ and ‘values.’ But these are not here used in a distinctively moral sense. Rather, they are used in the sense relevant to the agent, to indicate what he intends, likes, enjoys, or feels impelled toward and concerned about. It can certainly include his sense of obligation or of personal ideals. But we cannot take assertions about such things from a given individual as assuring acceptable status as moral judgment. Nothing about this is as yet moral, in the sense of having passed the relevant criteria for moral claims and judgments. In the case where the agent has what *he* takes to be a moral motivation, the decision-theoretic approach does not automatically endorse that agent’s claims. Instead, it subjects them to relevant tests. Our project is to find reasons why, in view of our situations and in particular our relations to other actors, we might want to invoke the kind of interpersonal judgments and rules claimed by morals to be applicable. The possible source of such reasons lies in analysis of where we would go, the two or more of us, if

people simply acted on their own independent values, wants, or reasons, and what, then, to do about it. Contemplating this may supply reasons—got from looking at things from an interpersonal point of view, as it were—for accepting various restrictions on what we might otherwise do. Those restrictions are the rules of morals. But they do not come from nowhere. The idea of decision-theoretically informed morals is to *derive* those restrictions from the various rational actions of individuals as they come into interaction with each other.

Two further words about morals are in order. There are various moral codes, sets of mores, in operation in various cultures and subcultural groups. There are also various moral theories. When morals is discussed here, our idea is to try to *construct* a rational view in the sense of a view that is rational for people to accept. It is not, then, to assume or presuppose some particular code of morals as being definitive of what morals is. If someone claims, for example, that in the theory partially expounded in this essay, certain things which he “knows” to be moral are said not to be part of rational morality, the question is whether there is some definite, fixed, known thing, ‘morality,’ such that it makes sense to say that that is what morality calls for even though we also know that it is not rational. It seems to me that a “morality” of which that can be true is useless, in a sense in which morality hardly can be. Morals purports to have authority over us, to tell us what to do in a way that brooks no compromise. Yet if it were irrational, whence then could come its authority over us presumptively rational beings?

Now, we do hope, but we also do not expect, that people will always do what they agree to be right. Moral rules, it is proposed here, are the ones that it is rational to subscribe to and uphold. They are not necessarily rules that one will always find it rational to conform to. We are tempted to do what is wrong, and there can well be a split between what we reckon to be right and what we want to do, and want it so much that we may overlook or repress our moral beliefs. However, we also know that the fact that we want to do it very much isn’t going to get us anywhere with those who are negatively impacted by our conduct: they cannot be expected to sympathize with us, in general. And that’s important. It is part of the very essence of morals that morally right action is defensible, wrong action not, and that those who do wrong are liable for correction at the hands of their fellows. Thus we have to do better than cite self-interest, when proposing defenses of our conduct to others.

It would be absurd to have a moral theory that “demonstrates” that people always do the right thing. Manifestly, they don’t. It is not absurd to have a moral theory which shows that what they do is the wrong thing, if it is, and why it is wrong, and does so in a way that the individual wrongdoer can see to be compelling. The hold of morals over us will be the hold of reason even if we can understand that it might, in a given case, not succeed in getting people to do the right thing.

To make sense of such possible disparities between actions and the rules which those very actors can see to be rational, we need to distinguish between two aspects of morals—distinguishable, but essential, and important. One aspect, of course, is *doing* the right thing—that is, in complying with the requirements called for by the best moral theory. But another aspect is what we may call the “administrative”: criticizing others, teaching people, including young people, and more generally attempting to influence the conduct of others. Moral rules and principles are a considerable, and fundamental, part of this. When someone queries a criticism, we cite a rule and suppose it to be a sound one—one that the accused will see to be sound.

Morals is interpersonal, and it is interpersonally based. It is so, because it has to be. If Jones tells Smith that Smith must do x, but all that x has going for it is something that seems sound to Jones, or appeals to him, or is to his advantage, why should Smith be

impressed? We must do better, and find a solution which makes sense from both points of view, given the facts of interaction. It may be that having done this, someone will persist nonetheless in doing what is contrary to their mutual interest. It is rational to think that mutually made agreements ought to be kept, and irrational to hold that it simply doesn't matter whether they are or not. That won't keep some people from reneging or simply violating those agreements. When that happens, we have problems, of a practical kind. We do not, however, have problems about the soundness of the morality in question. By contrast, a morality based on what makes no sense from the point of view of some who are nevertheless called upon to conform to it, *is* unsound, and stands as a view with no credentials. It, indeed, makes no sense. There are irrational moralities, in the anthropologist's sense of the term 'morality', that is, sets of rules that are professed by some. But to hold that morality is just like that is to raise the question, why bother? We can surely hope to do better.

2.3 Rational and Reasonable

John Rawls makes a now-familiar distinction, between the rational and the reasonable. "Reasonable persons, we say, are not moved by the general good as such but desire for its own sake a social world in which they, as free and equal, can cooperate with others on terms all can accept. They insist that reciprocity should hold within that world so that each benefits along with others.

"The rational is, however, a distinct idea from the reasonable and applies to a single, unified agent What rational agents lack is the particular form of moral sensibility that underlies the desire to engage in fair cooperation as such ... Rational agents approach being psychopathic when their interests are solely in benefit to themselves" (Rawls 1991, p. 50f.).

Rational persons are, in effect, Gauthieran "straightforward" maximizers; reasonable persons are Gauthieran "constrained" maximizers (Gauthier 1986, p. 167). The question, in effect, is whether reasonable persons *are rational*: that is, whether the rational person who is not reasonable as characterized is in fact being altogether rational.

Gauthier insists that "constrained maximization is not straightforward maximization in its most effective disguise. ... The constrained maximizer does not reason more effectively about how to maximize her utility, but reasons in a different way." (Gauthier 1986, p. 169f.). So the question is, can we show that it is rational to reason in this different way? Gauthier believes that we can. The rational agent "compares the expected utility of disposing himself to maximize utility given others' expected strategy choices, with the utility of disposing himself to co-operate with others in bringing about nearly fair and optimal outcomes." (Gauthier 1996, p. 171). He argues that, since the constrained maximizer—adopting the Hobbesian right of self-defense, in effect—will not cooperate with noncooperators and therefore will not be taken advantage of in those contexts, but will be able to cooperate with cooperators, thus securing himself the additional benefits made possible by cooperation, the constrained maximizer will expect to do better, in general.

We can hardly believe that he will always and necessarily do better in each and every particular case, however. Gauthier's further investigations involve making estimates of the likelihood that those with whom we interact are fellow cooperators or not. Rational agents cannot be assumed to be "transparent," but they will be to greater or lesser degree "translucent," in his nice terminology (Gauthier 1986, pp. 174–177 especially). Obviously, we are not perfect. We do not define the rational agent as possessing full knowledge, god-like. And a mistake could be made in a particular case, so that the other person does take advantage in that case. We trust the wrong people and we get taken. And in various other ways, imperfections of knowledge can lead to our undoing. Such is life.

Nevertheless, we can agree with Rawls that the rational but not reasonable person tends toward the psychopathic. A Platonic argument could be mounted, in behalf of the view that such persons lead a bad life, worse than awaits the nonpsychopath who is ready to engage in full cooperation with others. But this requires a type of argument not clearly available, if we accept that rational behavior proceeds from the basic values of the agent *whatever they are*. We might pity such agents, but it is not clear that we can claim to have demonstrated to them that they have made a mistake.

What remains is that the supposedly rational person who accepts no constraints deprives himself of any kind of support for his activities derivable from the acceptance of his behavior *by others*. Knowing this, the rational person will pay lip service to public morality. But in so doing, he of course lays himself open to the kinds of criticisms that we are familiar with in moral discourse. If he professes not to care about those, we can hope, and often realistically expect, that it will cost him dearly in terms that he does care about—because the rest of us will do our best to impose such costs on him in the event of noncompliance with the rules that he professes to accept. The rational individual making his way in the real social world that we all live in will learn to be reasonable, or face serious costs if he does not. It is not clear that we can do better than that, but also not clear that we need to. The rationality of morality lies in the expected benefits of erecting and successfully administering the institution of morality. Those benefits, in the general picture, are huge, as Hobbes pointed out. It is irrational to expect other people to put up with behavior that deviates substantially from the model of cooperative interaction, and it is irrational to assume that one can be continually successful in pursuing warlike policies against all peaceable persons. But it is, unfortunately, not impossible—Stalin lived to be 74.

3 Game Theory: Baselines and “State of Nature”

Game theoretic diagrams are usually presented with ordinal numbers: ‘1’ is best for the player in question, ‘4’ (in a 2×2 diagram) worst. Ordinal numbers make sense in relation to each other, but to interpret them in relation to actual situations we imply that there is some relevant “baseline” from which we depart in playing. ‘1’ is the maximally beneficial departure from this antecedent situation, ‘4’ the least beneficial such departure. But in fact, all could be negative departures—the whole game could leave us worse off, or alternatively, the least good outcome might be better than the situation *ex ante*.

How should we interpret this if we hope to use game theory as some kind of input to moral theory? There are two ideas, generally speaking. One is the classic idea of Hobbes and the other contractarians: the background is the “state of nature,” the state in which there are no moral restrictions on anything. The other alternative is the status quo *ex ante* whatever it may be. That is, the “game” is played in some particular situation with variable background, so that we are to think of the numbers as departures from the given status quo, *whatever it may be*.

- (1) The noncooperative outcome in prisoner’s dilemma for Hobbes is stark: life for all, in his oft-quoted words, “nasty, brutish, and short.’ On the other hand, moving to the $2/2$ cooperative outcome has all the lovely advantages of civilized society, with a measure of prosperity, knowledge, agriculture, and so forth. But some theorists imagine that the “state of nature” would instead be itself peaceable and cooperative.

Hobbes expounded his theory as having a specifically political output: the “sovereign” is what is needed to get us into the mutually beneficial state. What is in question here, however, is not the sovereign but *morals*—quite another matter. So we are to try to envision

the “state of nature” as the absence, not merely of government, but of *any moral sense whatever*. Hobbes’ dismal forecast for anarchy is, I think, very much more plausible if it is reconstructed in these terms. We can well imagine that persons living together in society but having no moral sense whatever, coupled with a likelihood of perceived gain from assorted activities that we all think to be evil, would indeed result in something like the “war of all against all” scenario, with attendant miseries.

- (2) If, alternatively, we apply prisoner’s dilemmas in the full variety that real life affords, the theoretical situation is wholly different. We might test a particular proposed moral rule of a fairly specific kind, for example: say, “always tell the truth.” But in the situation imagined in the famous letter to Kant, of a murderer coming to the door inquiring the whereabouts of his intended victim, which is known to the inquirer, we will surely get a different result than in a more normal situation of inquiring for some ordinary innocent but useful piece of information.

As efforts to add something to fundamental moral theory, the quasi-Hobbesian “moral state of nature”—the condition of “moral anarchy” as we may call it—is our default outcome. Here the social contract theorist’s claimed unanimity falls into place. What we all “agree” on is unlimited pursuit of our interests, whatever they are. The “immoralist” is in effect taken to be, by his example, advocating that kind of operation. So it will always be true, by default, that persons proposing to proceed by violence or chicanery can have no objection to being used in kind by any others. That is the basis for the second part of Hobbes’s First Law of Nature: if we can’t get peace, which is the absence of aggression, then we are morally entitled to use whatever means may be necessary to deal with the threats to which such aggressive behavior exposes us. Self-defense is a perfectly acceptable justification for the killing of a would-be murderer, for example.

3.1 Some Game-Theoretic Models

With these general caveats in mind, let’s look at some of the main interesting game-theoretically described situations. As usual, we use two-person “games” as our standard. An important question is how much and what kind of differences it might make to extend analysis to multi-person situations—but that is a question we will, for the most part, not address here, alas. That is because the essential move, from consideration in light exclusively of the agent’s motivations in which reasoning is, as David Gauthier puts it, *parametric*, to consideration in light of the fact that we are interacting with some other, independent agent, wherein we must perforce, if we have any sense, reason *strategically*. The two-person situation enables us to attend usefully to the difference that can make.

1. Prisoner’s Dilemma

The idea of decision and game theory is to enable results from mathematical methods. Decision theory is more general, dealing with all sorts of decisions, in principle; game theory concerns decisions in interaction with other players—usually just one other. Some of these are of very special interest—notably the Prisoner’s Dilemma, which dominates the literature (though there are others, some of which will get some attention below.) In PD there is partial conflict, and decisions can become very complex. Two players with two options each are so related that one player’s best outcome is the other’s worst, and vice versa, but they have a common second and third best. A familiar representation of this famous problem utilizes ordinal utilities, from 1 (“best”) to 4 (“worst”), the terms ‘best’ and

‘worst’ being applied on the basis of the particular agent’s own values. Thus the diagram looks like this:

		B	
		x	y
x		2, 2	4, 1
A			
	y	1, 4	3, 3

The common second (2, 2) is known as the Cooperative outcome, and the moves that lead to it are called “Cooperate”; if both attempt to get their best outcome, taking the option now known as “Defect,” they will in fact end up with their third best, whereas if both go for the second-best, both will get it. This is especially intriguing for moral philosophy, since it models the apparent paradox that in order to do our best in a social setting, we must deliberately *refrain* from choosing the “best.” Buying the goods and paying the asked price is, on the face of it, not as good for the agent as collecting the goods without paying, or getting the cash without supplying the goods. But of course, it’s only good for one agent, the other being exploited in the process. Most of us most of the time cooperate, and the long-run payoff from this general abstention from exploitative behavior is in general very good indeed. Yet it seems that one foregoes a possibly greater gain, in each case, by this behavior.

Now, if rationality is identified with picking one’s best option—as seems only reasonable—then morals is a problem. In the case of PD, would we cooperate? Not obviously. Yet cooperation is the obvious choice of morals. So, is morality rational, or isn’t it? That question has been the subject of an enormous amount of inquiry among philosophers. What are the options?

One idea is David Gauthier’s (Gauthier 1986 pp. 157–189) : it is rational to cooperate with others provided they will cooperate with you, and rational not to cooperate if they won’t. In each case, one deliberately chooses the second-best (so it might be argued); yet the *disposition* to cooperate is, he claims, rational. Gauthier proposes that the rational agent facing a Prisoner’s Dilemma, realizing that the other involved agent is in the same boat, will adopt as a *rational disposition* the disposition to cooperate with others who look as if they will reciprocate (Gauthier 1986, ch. VI, pp. 157–189). Of course, more precisely it is to cooperate with others who *will* reciprocate. Trouble is, you not only don’t *know* that they will, but you also know that if you act first, then it is apparently irrational for them to do so. Yet the exploiter—the “straightforward maximizer” in Gauthier’s terminology—will soon find others attempting to exploit him as well. The *conditional* disposition to cooperate, that is, to cooperate provided others do too, has a far higher payoff than the disposition to exploit, in any fairly longish run.

The Gauthier solution has a major problem, as the above suggests. If I decide to cooperate with cooperators, how do I know whether my interactee is a cooperator or not? There is no red letter ‘C’ on the foreheads of such people, after all (and even if there were—could we be sure it was correct, or is he just cleverly trying to mislead me?) It seems we would have to be

guided by his past behavior. Yet Gauthier claims validity for the one-shot PD. Fortunes could be lost by persons trusting to this view, it seems. Gauthier provides an account of “transparency” and “translucency” in which we make judgments about the likelihood that the individual we interact with is a “constrained” rather than a “straight” maximizer. (The latter are paradigmatically ones who would default on a prisoner’s dilemma.)

Much more popular among game theorists than Gauthier’s idea is to appeal to iteration. Suppose we play prisoner’s dilemma over and over with the same party, and also that we don’t know which is the final round. It appears we will then do better to play tit-for-tat (the celebrated strategy in which we cooperate on the first move, and then do whatever the other party does in response.) As indeed we will, if both play it and follow that rule—but better than what? Not, perhaps, better than someone who occasionally defects, this appearing perhaps to be accidental so that I try to resume profitable interaction after awhile. And indeed, Russell Hardin has plausibly argued that there is no decisive theory for prisoner’s dilemma (Hardin 2003). Game theory is indeterminate about this.

Still, if it is repeated often and with no one able to know whether this is the last iteration, then it becomes plausible to say that rationality generally recommends cooperation. Or at least, plausible on certain general assumptions, especially the assumption of symmetry, according to which both players, being rational, will play the same strategy ... which is where the fun begins. The moralist has an easy victory if that assumption is allowed. But why would it be?¹

At this point, we are at least tempted to bring things into the picture that pure game theory doesn’t obviously welcome. Do we cooperators get pleasure, as well as benefit, from cooperation? Do we simply *like* to cooperate? If so, does this affect the decision situation? It surely would, after all. But how? Do we change the numbers in our decision-chart in order to reflect these extra “payoffs”? That would certainly complicate decision-theoretic life. Most plausibly, it would have the result that we aren’t, after all, really playing Prisoner’s Dilemma. Perhaps we began doing so, but then the game subtly changed as we went along. Or perhaps we never were in the first place—we aren’t, as it were, built in such a way as to play the game in pure form.

This last remark leads us to what is, in one respect, perhaps the most fundamental problem with game-theoretic analysis in application to real life: namely, that life, and therefore decisions, are complex, and made in very rich environments, so that what is entered on our decision diagrams is bound to be less than all that is relevant. Or if we simply stipulate that the diagram entries are to be understood as having somehow taken into account everything relevant, then we will never know whether our diagram is right or not. We will, therefore, never be able to be sure that we in fact are in such a dilemma, speaking absolutely. Stipulation is no substitute for facts. The point, then, is that charts are incapable of capturing the “everything” that must be “equal” in order for the results in the diagrams to reflect reality. Life is complicated, and we rarely know our own all-up values well enough to be sure that a decision diagram captures the situation we regard ourselves as being really in. (The foregoing point may be taken to be a special case of this: a taste for cooperation is an “other thing” relative to the desirability of my having your Mercedes-Benz.) Thus even if our reasoning is perfectly valid, there is an inevitable problem of soundness: in effect, the premises of game-theoretic arguments can scarcely ever be known to be true, insofar as they meant to describe what actual people do.

Thus, for example, consider the well-known results of empirical studies of Prisoner’s Dilemma. Not only do various samples vary enormously between selection of the “defect” (uncooperative) and the “cooperate” strategy, but also we are told that economics students consistently play the “defect” strategy much more often than other students, as well as that

¹ The issue is explored, somewhat infuriatingly, by Ken Binmore (1994), chs. and 3.

they play it at the end of the term much more readily than at the beginning. The natural hypothesis to form about this is that most ordinary people we know are pretty cooperative, and studying game theory tends to make people less so. Does that speak well for economics students? And badly for ordinary people? Hardly! Does it mean that ordinary people attach some extra utility to cooperation? Maybe. Does it mean that life for ordinary people goes better? Yes. And game theory *does* teach us that, or so I suggest.

Provisionally, the point is this. Despite the indeterminateness of prisoner's dilemma, real people in society can see that cooperation is what *we ought to* direct our social pressures at. That is the stuff of social rules. As a member of society, realizing that we are often in such situations, no rule other than cooperation makes any sense, if intended for general inculcation and direction to all. For after all, why would the person who would be exploited accept such a rule? If we ask, why would the person who would win accept it?—the answer is easy: it's not *his* rule, it's *our* rule. And if we know that people will be like that, they won't be able to "win"—or at least, it will be rational for us to attempt to prevent them from doing so. Moreover, everyone knows that they are just as eligible to be the victims as the victors in such dilemmatic games. We have reason to react against exploitation. Cooperation is *optimal*: best for each, so long as others are equally capable of being uncooperative, to our disadvantage.

Note that the point here is that calling for cooperation, across the board—barring, as always, special cases in which more things are unequal even than usual—is rational, *not* because cooperation is itself uniquely rational for each player on each occasion. Rather, what's rational is to have a general social rule of cooperation. Socially speaking, cooperation is peaceable and productive; noncooperation motivates people to get into fights, and in general relate in unproductive and unhelpful ways. What is best for each is to accept a rule of cooperation, and then to abide by it, given that others will too, of course.

Prisoner's Dilemma is far from the only game-theoretic interaction of interest. Four others (at least) deserve notice here: zero-sum, pure coordination, battle-of-the-sexes, and Chicken. I comment on them in that order.

2. Zero-Sum Games

The defining feature of the so-called "zero sum" is that one player's gains exactly equal the other player's loss: where A and B are the players, A gains if and only if B loses. The changes from the starting point sum to zero. Player A's interests—within the game—are absolutely incompatible with player B's. The diagram looks like this:

		B	
		x	y
A	x	1, 0	0, 1
	y	0, 1	1, 0

If we leave it at that, it is obvious that there is no such thing as a *social* “solution” to zero sum games—no basis for a uniform rule, acceptable to all, to the effect that A, or B, should win. (It is irrelevant here to point out that the theory of zero-sum games is complete and calls always for the same strategy. That is the rule for individuals, but it is not a rule prescribing a socially enforceable solution to zero-sum problems. A rule put out for the benefit of everyone in society cannot exist in such a situation; it will produce no such benefit, since any benefit for A is a detriment for someone else, B. Put abstractly, the situation of total conflict can have no mutually acceptable outcome.)

But again, there is the problem of what the numbers actually mean. For example, most games in the usual sense of the term are competitive: A wins if and only if B loses. This raises a question: why do people ever play them? In fact, they play them a great deal—all sorts of competitive games are among the commonest pastimes of human beings. If game theory were to be understood to make it puzzling why people do this, one would think, then so much the worse for game theory! But to say that is to be indiscriminating. Zero-sum games being zero-sum does not mean that they are overall, dead losses for the loser in terms of his overall set of interests. Quite the contrary: we *like* to play (that is, lots of people do, in various specific games—and also, to be sure, lots of us don’t. And hardly anyone one likes to play just any and every such game.) Liking to play means that *we view ourselves as being better off playing, even if we lose, than if we don’t play at all*. One way to represent this is as follows: suppose that the pleasure of the game is +8, while the maximum loss, if we lose, is -4, and the maximum gain, if we win, +4. Then each player gets a net gain of at least 4 from playing rather than not playing. And, of course, for those who don’t like to play, the decision is simple: don’t play! Since that is virtually always an available option, there is no problem. The *moral* upshot is easy: be sure that any zero-sum games are mutually voluntary.

The point is, then, that most of the things we call zero-sum games are embedded in larger contexts such that “play!” (and “play your best!”) is a good solution for both parties: not because they will both win, which is logically impossible, nor because they will “share equally” as if all games would be ties, but because win or lose, to play is more beneficial—more interesting, for instance—than not to play, even if one loses. Meanwhile, it is voluntary. One always has the option not to play, when the subject is games of the usual sort. So those who would not find it interesting or rewarding simply need not play, no problem. And all that is something a game theory diagram will *not* teach us.

On the other hand, however, when choosing not to play is *not* an option, then the very fact of zero-sum means that we have terrible problems, since it is by definition impossible to satisfy everyone in a fully competitive situation. Social philosophers often depict, say, the situation of resource ownership in the world as a zero-sum game. Since there is only a finite amount of coal or whatever, more for A is less for B—so it is said. But this is undiscerning. To begin with, the finite amount of coal that there is known to be is vast, and more for A within the next several hundred years, say, does *not* mean less for B. Moreover, it may well mean *more* for B: A’s having coal *now* may enable A to increase the production of future coal, more of which finds its way into the furnace of B than would otherwise have been the case; or it enables him or them to pursue research, in nice warm rooms, leading to alternative energy sources in the further future. We have to be *very* careful about drawing conclusions from the general fact of scarcity as if it implies a zero-sum situation, even when, in some sense, scarcity *is* a general fact. Just as with games, which we enjoy despite their zero-sum character, so with resources of many

kinds: if the zero-sum structure operates against a background of productivity or abundance, it is not necessarily a problem. The enormous abundance for all in wealthy societies is not due to their successful exploitation of supplies at the expense of somebody else.

On the other hand, if it really were the case that more for A necessarily means less for B, in local cases, then society *would* have a problem—as Hume pointed out in his famous thought-experiment about utter scarcity (Hume 1751, III.1). As an example, consider malice. If A is malicious in relation to B, then A “loses” when B gains, and vice versa. The social solution here is to bring up young A in such a way as to eliminate or greatly contain his level of malice. Malice, envy, a fixation on social status and various other attitudes are vices, precisely because they are, in ways that are meaningful in real-world contexts, zero-sum. The moral for society is to suppress such attitudes, labeling them as vices (which they are), meanwhile allowing hockey, bridge-playing, and other generally amiable, even though zero-sum, games to flourish.

A further moral about zero-sum games of the familiar kind is that they are framed by their rules, which define the game and which also limit the zero-sumness of the game to very restricted ranges of values. The intrinsic value of getting a small rubber disc into a net is approximately nil for almost everyone. Getting it through that net against determined opposition more often than the other team gets it through one’s own is another matter altogether. The rewards, either intrinsic or in the way of audience approval and, in the case of professional players, salary as well as general glory, are potentially high. This tempts some to cheat—getting those goals outside the restrictions imposed by the game-defining rules. The effect of cheating is to lower the level of socially compatible utility from game-playing, setting some against others in entirely avoidable ways. Again, a game-theoretic chart used to help a team with its strategies, say, will not take such things into account. The confinement of zero-sum utilities to a very narrow range, and their embedding in a wider context that is fully cooperative, is the secret of success for competitive games.

At a very different level, consider “warrior cultures.” Why do some people seem to like to fight, even at the level of, say, dueling, or war? Notoriously, the warlike temperament is a frequent feature of the human scene. Wars have high costs for both winners and losers: many on both sides die, property damage in the terrain of war is extreme, and costs in toil and foregone leisure even for the “winners” are enormous. If a given group or nation is attacked, it is not surprising that it will strive to defend itself if it can. Less obvious is why the attackers attack. But if it is perceived that the possible gains from war exceed the known costs, then also the nation or the culture that develops that perception will also likely cultivate the warlike temperament. Spartan youths, we are told, spent most of their waking hours training for war, Spartan mothers were encouraged to feel immense pride in their fallen sons, and so on. Not surprisingly, Sparta was a poverty-afflicted place, and the average expected life span pretty short. The sober philosopher contemplating all this is surely on a plausible track in concluding that war is to be generally deplored rather than encouraged, and to hope that the world at large can find its way to “perpetual peace” (as in Kant’s famous essay) rather than frequent war. It would be one thing if we were all Spartans: war might then be viewed in something like the way in which participation in soccer or hockey is now. But we aren’t—most of us—and a social rule enjoining us all to devote ourselves to the martial virtues would be absurd as a general rule for all. The rule of seeking peace, on the other hand, has precisely that status. As a general rule for all, it is exactly right.

3. Pure Coordination (ordinal utilities)

		B	
		x	y
	x	0, 0	1, 1
A			
	y	1, 1	0, 0

We have a coordination *problem* when, if all do x, all gain, and if all do y, then all likewise gain, and equally, as compared with all doing x—thus giving the individual no reason, on his own account, to choose x over y or vice versa. So now—which will we do? In the absence of effective communication, if we attempt to coordinate by individually flipping coins, we will end up with very ineffectual coordination—for a two-person situation, 0.5 is the expected payoff rather than the 1.0 in principle available for all. What to do? Social philosophers have gone so far as to make this an argument for monarchy, or at least for government. But that’s too quick. There are other ways to coordinate. For one thing, we can select an arbiter just for this occasion or this set of occasions, leaving other cases to others. Another aspect of coordination is that *custom* is coordinative: everybody follows the local custom just because others do, and in the many cases where there is no real gain from having custom x rather than custom y, the gains from having some custom or other are large. When, on the other hand, there is no established practice and there are gains to be made from having one, efforts to achieve coordination are well taken, and because the aim is desirable for all, have a good chance to be effective. A useful moral rule, then, is “in the absence of special reasons to suppose that some other custom would work more good or less harm, do the done thing!” Good manners, for example, are generally to be recommended, and they take their specific character from the “done thing” in that society or in that circle of society.

While all doing x is neither better nor worse than all doing y, it may be that a would-be coordinator would much prefer x to y, for reasons of his or her own. And realistically, in many-person cases it is quite unlikely that all doing x really will be equal to all doing y; also likely is that if the coordination is among *almost* all, that will be good enough to do, and still worth achieving. And if what is claimed to be a coordination situation requires coercion to achieve it, then what is achieved is not truly coordination, we may well suspect. This segues into another sort of game—one of the two following

4. “Battle of the Sexes”

This is a case of less than perfect, or partial, coordination. A and B have different tastes: A likes to do x, B to do y; still, A would rather do y with B than do x alone; and B would

rather do x with A than y alone. Worst for A would be A's doing y alone, and for B, B's doing x alone.

		B	
		x	y
	x	1, 2	3, 3
		A	
	y	4, 4	2, 1

In "Battle," like Prisoner's Dilemma, there is a common interest and a disparate interest, with an arguably (and normally) optimal solution available. Clearly we want to end up with either x, x or y, y—but which? One possibility is that we always use x, x, say because the man compels the wife in that direction; or vice-versa—the man is beaten into submission. (However, either of those would get us into a different game.) But *society*, surely, will recommend a fair solution: randomize between the two, often at 50–50. (There is the additional possibility that A's preference is stronger than B's. Then a problem is that A will always get his (or her) way. But an interpersonally rational solution here is for both parties to perform a weighted randomization, so that both sometimes do x, sometimes y, with one predominating in proportion to the relative intensities of preference. Indeed, in many marriages, just that is the practice.)

Here again, concentration on just one fairly specific context would be a mistake. Possibly there is another context, again with unequal preferences, but such that coordinating always or mostly on x, x for the one context, and on y, y for the other, gives each party the optimum result. (This might be some people's argument for slavery, in some circumstances. Somebody has to do the rough work; but if only rough work is done, that would be bad for all. But if only arts and letters are done, we all starve. Both, says the defender of slavery in these circumstances, do better if there's a slave class and a master class. But then, it would surely be still better for all to achieve the desired division of labor without anyone's enslaving anyone, if possible. And whatever might conceivably have been the case in some societies at some times, it is clearly possible now; and so, slavery now is universally, and rightly, deplored.)

5. Chicken

In this particularly vicious game, there is a common *worst*.

		B	
		x	y
	x	2, 2	1, 3
		A	
	y	3, 1	4, 4

The classic example, from the movies, has two cars full of teenagers driving directly toward each other; the one that swerves first is the “chicken.” But if no one swerves, they have a head-on collision and all get killed. The “cooperative” option has them either not playing at all, or both swerving simultaneously at the last moment. The atomic arms race provided a more serious classic example. A threatens to “nuke” B unless B backs down on some point; B threatens to do the same to A in retaliation or in order to induce A to back down on some other point. If neither backs down, there’s a nuclear war. The cooperative option has both sides backing down, or mutually disarming, or just mutually deciding not to play. The latter in fact obtained. The nuclear standoff, though, more nearly exemplified PD, with the superpowers standing off at 3, 3—and wasting enormous amounts of money on an option dominated by mutual peace.

Success at chicken is a function of the plausibility with which the player can threaten the defect (“hawk”) option. If both players are really stubborn, disaster looms for both. They’d be better off with any other combination than y, y . But which? A wants y, x ; B wants x, y . Each can avoid y, y by conceding to the other party; each would rather that the other concede. What to do?

But it is perfectly clear what *society* should do: discourage chicken, and encourage the cooperative option. The problem is—how is it to do this? And what, even, does it mean to say this—seeing that all decisions are made by individual people? An additional large problem is that threatening may just get it into another game of Chicken with the parties it hopes to induce to cooperate.

A further point about chicken is that this game makes it obvious that the absolute size of the utilities involved is pretty important. The State compels us all to pay taxes, which most of us would rather not. If I resist and it prosecutes, we both are worse off than if I just knuckle under or the State doesn’t press its claims. But I am ever so much worse off than the State in the 4,4 option. Almost all of us knuckle under. If the payoffs in the 4,4 position were more nearly equal, then the State might not have such an easy time of it. (In a state of civil war, for example, there are large losses for both sides; both have reason to find a compromise.)

I have spoken above as if we had our choice about whether to get into a game of chicken (or any of the foregoing.) But the philosopher might argue that we have no choice, and that our natures or situations, or some such thing, condemns us. I think we can plausibly put this down as irrationalist, but will let the matter rest here, other than to say that if we allow the 2,2 outcome to be an available option, then it doesn’t matter whether the game is inevitable, especially if it is reiterable. We can see what’s coming, and choose to “cool it,” and especially, choose to advise all to cool it, in general, confining their “chicken” interactions to minor, trivial matters.

4 Morals and Games: Social Rules and the Individual

To return to the question not, perhaps, obviously answered in the above discussions I say that “society” should discourage Chicken games, but since we are all individuals rather than “societies”, what sense does that make? There is, in fact, a good answer to this. Individuals are actors in game theory, but they are also students of game theory, for example. We can see what’s coming if people get into (serious) chicken games, and *we*—meaning, anybody—*can* exercise such powers as we have to attempt to head people off from doing so. Interestingly enough, the threatening of evil consequences for noncompliance is itself, usually, an example of chicken at work. However, it is often one where the 4, 4 outcomes

are very different for the two “players.” An organized force of people threatening an individual has a good prospect of success, because what the individual has to lose if he plays Hawk is far greater than what the many members of the group have, individually, to lose if he does.

That of course generates further concern about the powers of groups, which are enormous in a great many cases. But there are groups and groups, and if some group exercises its powers over small groups or individuals in certain ways, the rest of us can see the point of being concerned about the large group getting out of hand. I would argue that there is a general rule on which all will do best to settle over the very long run and the very widest “group”—the group of mankind generally. That rule is the familiar non-harm rule, which has a long and honorable history. We need much more than just game theory to get us down that road—but game theory helps enormously to see what the problem is and how that rule is a good solution.

Is morality rational? That question is usually construed in such a way as to imply that the answer is affirmative only if each and every action called for by a rational morality is itself rational in the sense of utility-maximizing for the agent. That is clearly an impossible requirement. To appreciate its impossibility, we have but to recall that morality asks us to curb our passions, to give way, sometimes, to others, to forego the possible payoffs of doing evil to others. Anything claiming to be a ‘morality’ for all (as opposed to the rule for some narrow group, such as the Mafia) would be unrecognizable if it did not have those features. What the decision-theoretic moralist must be saying is, instead, that morality pays off, somehow, in a different way. The obvious way would be that even though every now and then we settle for less, we very often get more than we otherwise might, and from our own point of view, the longer-run gains outweigh the losses.

In what does the thesis that morality is founded on rationality consist, then? The answer is to be got by attending to the standard, traditional foci of what we might call the “moral project.” Aristotle held that moral virtue consisted in a disposition to strike the “mean” in matters of passion. Reason is to control the passions, and to do so by somehow being able to size up situations to see when we have gone too far or not far enough. But Aristotle’s morality is much wider than what is being discussed here. And it has its apparent oddities. Why should we refrain from murder? The suggestion that it’s because if we do, then the passions which might incline us to murder will exemplify the right “amount” on some kind of internally constructible scale seems downright bizarre. Plainly, what’s wrong with murder is that its victims end up dead, which (we assume) they don’t want to be. Cooperation doesn’t sanction murder. Whatever the gains to A from murdering B, morality holds that they cannot justify that particular method of acquiring those gains. And similarly with all of the other ways in which we might hope to promote our own good by imposing injury and damage on others.

Consider now the Prisoner’s Dilemma. If I propose to take the Defect option, you are motivated to do likewise, leaving us both worse off than if we had both chosen cooperation in the first place. The moral person does not just test the waters to see whether his intended victims will in fact take that option. Instead, he or she sets, it is hoped, an example for others to do likewise. Attending to social iteration, we can argue that the longer the run, the greater the benefit to all from all having a preference, and an ingrained tendency, to cooperate. What makes this plausible—insofar as it is? I suggest that it is, above all, the relative plausibility of Thomas Hobbes’s (Hobbes 1651, Ch. XIII; 1950, p. 101) thesis that people are approximately, broadly, equal, in respect of their capacity to work ill for others. In particular, Hobbes asserts, “As to strength of body, the weakest have enough to kill the strongest.” If virtually anyone has the capability of killing virtually anyone, then we will all

do well to subscribe to a moral regime of classifying murder as something not to be done, and something to go fairly far out of one's way to try to help bring it about that others don't either. By contrast, consider David Hume's point about animals: they simply lack the capacity to impose substantial, premeditated dangers to people (Hume 1751, sect. III, part I).

Some of them do have the capacity to kill us, but they lack the resources, generally speaking, to be trained from early on to refrain from utilizing those capacities. We humans can't deal with animals in the moral way. We can fence them in, kill them, turn them into hamburgers, or, often, into domestic pets, yes; but we can't enter into a sort of "social contract" such as we decision theorists claim morality among humans to be. Animals simply are not our equals in the ways that would enable us to benefit from entering into serious moral relations with them. (For all that, plenty of pet owners, or keepers of domestic animals, or even zookeepers, are ready to speak up for the animals. But the scope of such appeals is extremely limited. Meat-eaters, for example, do not reasonably expect a positive return from granting animals the right to life.)

The relevant feature of contract, as Hobbes pointed out, is that one of the two parties usually puts himself at the mercy of the other party for some period of time. If I pay now and you deliver tomorrow, what's to keep you from failing to deliver and pocketing my money anyway? There are two kinds of answers to this, both relevant. One, of course, is that you'd like to stay in business, and the contribution to your bad reputation if you don't deliver may be enough to make you regret your noncooperative participation. But the other is, simply, morality: that is, an inner aversion to such behavior. You can and usually do get such things from childhood training, peer pressure, and general socialization. Game theory makes one appreciate why such institutions are a good idea. The benefits of being in a regime of cooperation are enormous, but they are only available to persons who do not try to maximize against others on each and every occasion of their lives. We will all do better if we do not always think about doing better—especially, about doing better than the other fellow.

It is game theory itself that paves the way to such conclusions. On the one hand, a morality of nothing but self-sacrifice is a nonstarter, to rational beings. But on the other, a morality enabling one to capture the benefits it makes available if done right is a very good investment, so long as we are among enough others who see it that way to make it so. We are all in some degree dependent on each others' good will. Because we are, having good will ourselves is the most plausible choice.

We have to add a serious cautionary note to all this. Hobbes' postulate of equality, as we might call it, is at best an approximation, and at worst a serious distortion. For almost any persons A and B, A *can* kill B, and vice versa, yes: but for many pairs at many times, it is very much easier for A to kill B than vice versa. Moreover, persons banding into smallish groups to advance their causes can do so, very often, quite successfully in considerable disregard of the supposed requirements of morality. Genghis Khan, we might think, did pretty well. What can perhaps be suggested is that it is irrational for anyone, large or small, to *claim* that he is in a position to disregard morality altogether. It does not follow that it will always *be* rational to act accordingly. Doing so will earn you a lot of negative input from a lot of people, but for all that, moral considerations may not be enough to divert the agent from some harmful path. To have established that interpersonal cooperation has a great deal to be said for it, and that it is to one's benefit to claim that one is all for it, is not to have established that people will always, or in some cases perhaps ever, pay much real attention to it. The rest of us have to take hard lessons from that fact, but precisely how to deal with the noncompliers in all cases, lies rather beyond the terms of reference of this exploration. The world we live in is not always a very nice place. On the other hand,

cooperation can make it a great deal nicer, and that includes cooperation in the business of curbing the malevolent tendencies of the not-so-nice. Misplaced cooperation, on the other hand—cooperation with non-cooperators—is an invitation to disaster.²

To ask whether morality is confirmed by game theory is to ask an ambiguous question, at the least. If we are arguing against a background status quo of just any sort whatever, and ask whether in *that* situation the cooperative outcome is the maximal one for any given agent, then it is essentially self-evident that the answer is No. But if the question is whether we will do better to propose, attempt to instill in others, and generally to support a rule calling for cooperation, the answer, I think, is plausibly in the affirmative.

Morals is for us all, for people in general. All can and do engage in interpersonal encounters with partial structures of the sorts examined in game theory. The importance of the project of formulating general rules emerging from such contemplation is that if we play our cards right, we stand to gain. We don't play them right by ignoring others, but rather by establishing general rights for all, and doing what will promote general respect for those rights. That, I think, is a clear moral of game theory, examined not for its own sake alone but for supplying useful insight into the human condition, and useful remedies for what that enables us to see can easily go wrong. Game theory is not morality as such; the rational agent maximizing his utility may do so in ways we should not approve. And he should pay attention to our disapproval, though we know that sometimes he won't. Nevertheless, the odds are on the side of cooperation in society. Game theory itself enables us to see why it makes sense to erect social barriers against some sorts of actions, and socially administrable devices to promote others. Nothing is perfect, but it is perfectly rational to try.

References

- Aristotle (1941) *Nicomachean ethics*. Random House, Basic Works of Aristotle, New York
- Binmore K (1994) *Game theory and the social contract*, vol. I. MIT
- Gauthier D (1986) *Morals by agreement*. Oxford University Press, Oxford
- Hardin R (2003) *Indeterminacy and society*. Princeton University Press
- Hobbes T (1651/1950) *Leviathan*. Everyman Library, New York: E.P. Dutton and Company.
- Hume D (1751/1957) *Inquiry concerning the principles of morals*. Bobbs-Merrill, Library of Liberal Arts—Indianapolis
- Jeske D (2008) *Rationality and moral theory*. Routledge, New York
- Kant I (1797/1965) *Rechtslehre*. *Metaphysic of morals*, Part I, I. C. (Trans: Ladd J). Bobbs-Merrill, Library of Liberal Arts—Indianapolis
- Locke J (1690) *Second treatise of government*. Cambridge University Press, *Two Treatises of Government*, Cambridge 1965
- Narveson J (2006) Is pacifism self-refuting? In: Bleisch B, Strub J-D (eds) *Pazifismus*. Haupt, Ideengeschichte, Theorie und Praxis, Bern/Stuttgart/Wien, pp 127–144
- Rawls J (1971) *A theory of justice*. Harvard University Press, Cambridge
- Rawls J (1991) *Political liberalism*. Columbia University Press, New York
- Rousseau J-J (1762/1978) *The social contract*, chapter III. In: Masters RD (ed) *Tr. Judith R. Masters*. St. Martin, New York

² Further discussed in Jan Narveson (2006).