Thinking How to Live with Each Other

ALLAN GIBBARD

The Tanner Lectures on Human Values

Delivered at

The University of California, Berkeley
February 28–March 2, 2006

Allan Gibbard is Richard Brandt Distinguished University Professor of Philosophy at the University of Michigan. He graduated from Swarthmore College and received his Ph.D. from Harvard University. He taught at Achimota School in Ghana while in the Peace Corps, and at the University of Chicago and the University of Pittsburgh. He has also been a visiting professor at Stanford University. He is a member of the American Philosophical Society and a fellow of the American Academy of Arts and Sciences. He has served as president of the Central Division of the American Philosophical Society. His many publications include *Manipulation of Voting Schemes: A General Result* (1973); *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (1990); *Moral Discourse and Practice* (coeditor with Stephen Darwall and Peter Railton, 1997); and *Thinking How to Live* (2003).

I. INSIGHT, CONSISTENCY, AND PLANS FOR LIVING

Jonathan Haidt, the social psychologist, entitles a fascinating article "The Emotional Dog and Its Rational Tail." His topic is moral judgment, and the emotional dog is what he calls "intuition." Mostly, he argues, we don't arrive at our moral conclusions by reasoning. We jump to them with emotional judgments, with "affectively valenced intuitions," as he puts it. We will often be firmly convinced that our moral judgments rest on sound reasoning, and that unless others are driven by bias, they will appreciate the force of our arguments. He calls this the "wag-the-other-dog's tail" illusion. In fact, though, in our moral reasoning, we are not so much like intuitive scientists following the considerations where they lead, but like intuitive lawyers, reasoning to preordained conclusions. Reasoning is effective on occasion, he concedes, with "adequate time and processing capacity, a motivation to be accurate, no a priori judgment to defend and justify, and when no relatedness or coherence motivations are triggered." Mostly, though, what reasoning does is to construct "justifications of intuitive judgments, causing the illusion of objective reasoning."[1]

All this chimes in with Hume's dictum, "Reason is, and ought only to be, the slave of the passions." Haidt himself isn't talking about how moral judgment *ought* to work; he is offering a psychological account of how moral judgment *does* work. Now, even philosophers who stress reasoning have often thought that reasoning must rest ultimately on intuition. Intuitions give us the starting points of reasoning, and they tell us what follows immediately from what. Reasoning thus strings together a series of intuitions. Haidt's thesis isn't just that intuition is crucial to moral judgment but that it isn't this stringing together that mostly drives moral judgment. Reasoning he defines as going by conscious steps, so that it "is intentional, effortful, and controllable and that the reasoner is aware that it is going on."[2] What's powerful in moral judgment, Haidt argues, will be the single, emotionally valenced intuition that reaches its conclusion all by itself. Moral judgment doesn't have to be this way, for all Hume's dictum tells us, but that, Haidt argues, is the way moral judgments mostly are.

[1] Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgments," 822. Full source information pertaining to notes is provided in the bibliography at the end of the lecture.

[2] Ibid., 818.

We can ask whether what Haidt means by "intuition" is what philosophers have traditionally meant. The paradigm of intuition in the philosophical tradition has often been the geometric intuition by which we know the axioms of geometry. These are emotionally cool, whereas the intuitions that drive most moral judgment, according to Haidt, have emotions attached. It's an important question whether the intuitions that ground morality are always tied in with emotion, but that's not a question I'll be addressing. Later on, I'll distinguish senses of the term "intuition," but I won't restrict the term either to "hot" or to "cool" states of mind.

Now, we philosophers aren't expert psychologists. It's how we *ought* to reason that we are specially charged with assessing. Often we do reason, even on moral questions, and I'll assume that sometimes we should. The philosopher's job in particular is to reason, and if we ought never to reason on morals, then we moral philosophers may need to find another line of work. In this lecture, though, I won't engage in moral reasoning; that is for the next two lectures. My questions in this lecture will be *about* moral reasoning. What is its subject matter? I'll ask. How ought we to reason? If reasoning strings together intuitions, why trust its intuitive starting points? I'll talk about these broad questions in this lecture, and then in the next two scrutinize a particular piece of moral reasoning, one that purports to get remarkably strong moral conclusions from plain and clear intuitions.

Moral intuitions are puzzling. We seem to have moral knowledge; indeed, some moral truths seem so utterly clear as to be pointless to state. It's wrong to torture people for fun. Other moral questions are agonizing to ponder. Are there any conceivable circumstances in which we would be morally justified in torturing someone? If we have moral knowledge at all, it seems this knowledge must rest in the end on powers to intuit moral truths. G. E. Moore a hundred years ago elaborated arguments that moral claims aren't claims that could be brought within the purview of natural science. Two people could agree on all the facts of empirical science and still disagree morally. They could disagree, say, on whether, as Henry Sidgwick thought, pleasure is the only thing worth wanting for its own sake. The fault of the one who is wrong needn't rest on ignorance of the facts of nature, or failure to grasp the concepts involved, or any failure of logic.[3] Natural facts and conceptual truths aren't enough to entail answers to moral questions. If we are to have any moral knowledge at all, then, the gap must somehow be filled. What else could it be filled by but a power of

[3] Moore, *Principia Ethica*. The argument of Moore's that I find powerful is the one on p. 11 that I call his "What's at issue?" argument.

intuition, a power to apprehend some basic moral truths though not by the senses?[4]

Not all philosophers agree that morality lies outside the scope of empirical science, but I'll be offering a picture on which it does, and proceed on the assumption that the picture is right. Moreover, I would argue that even claims about rationality in science aren't entirely within the subject matter of science. Science itself rests on intuitions about the justification of empirical conclusions. If that's right, then it may not only be morality that raises puzzles about intuition.

In the case of morality in particular, a chief puzzle is that it is hard to see how beings like us could have powers of moral intuition. We are parts of the natural world. Crucial aspects of any moral truth, though, don't lie in the natural world. When we look at ourselves as parts of the natural world—as Haidt does—we won't find a responsiveness to anything nonnatural. We won't even find the purported facts we claim to intuit.

I'll begin what I have to say by sketching a view of ourselves as a part of nature. Moral right and wrong form no part of this view. It is part of the view, though, that we would ask ourselves moral questions, and come to conclusions about them. How things stand morally is not a part of the natural world, but our study of these matters is. (Later I'll be qualifying this, but for now let's stick with it.) Beings who think and reason about what to do, I say, answer questions of ought, at least implicitly, when they settle on what to do. Beings with our own psychic makeup make specifically moral claims. I'll speculate how these activities look within a value-free scientific picture. After that, I'll turn to the plight of the beings like us who figure in the picture, beings who think about what to do and think about right and wrong. Our answers to the questions we address will rest on intuitions—but, I'll be asking, if intuitions are the sorts of states that figure in Haidt's picture, why place any stock in them?[5]

Nature, Oughts, and Plans

Begin, then, with us as living organisms who are part of the world of living organisms. The upshot of natural selection is that genes become amazingly good at, as it were, working together to use us to make more copies of themselves. How, a great puzzle runs, have metaphorically selfish genes

[4] Sidgwick, *The Methods of Ethics,* 338–42, argues that ethics requires at least one intuition.

[5] The picture I develop is given in my books *Wise Choices, Apt Feelings: A Theory of Normative Judgment* and *Thinking How to Live.* For a discussion centered on intuition, see my "Knowing What to Do, Seeing What to Do," and for second thoughts on the theory of moral concepts in *Wise Choices,* see my "Moral Feelings and Moral Concepts."

come to make people who are, literally, not entirely selfish? The answer can only be a long and controversial story, and I won't address this particular story in these lectures. Rather, I'll ask about the place of *ought*s in the story, in this world of *is*s.

The story proceeds in terms of the metaphorical interests of the genes, the things that promoted their multiplying as the human genotype was formed, and, on the other hand, the literal aims, beliefs, and feelings of humans and protohumans. Genes proliferate in part by forming recipes for organisms that keep track of the world around them, very much including the social world.[6] Knowledge guides action. But it guides action in ways that proliferate genes only if the actors have the right aims, the right propensities to use their knowledge to guide action. Knowing where a lion is doesn't promote one's genes' reproduction if one's reaction is to try to pet it.

The beings in this biological picture of us, then, face questions of how things are, but those aren't the primary questions they face. The primary questions are ones of what to do, of what to aim for and how. Most organisms of course can't be interpreted, in any full-bodied sense, as addressing these questions and accepting or rejecting answers to them. Dogs chase squirrels and bark at intruders, and much of the time, we, like the dog, just act out of habit or emotion. We, though, of an intricately social species with language, differ from other animals in two important ways. First, our social emotions are especially refined and elaborate. A substantial amount of the human neocortex seems to function in the workings of emotions, and emotions include impulses to action. Many of our feelings are intensely social, as with guilt and resentment, with shame and disdain. Second, we are beings with language, and we make judgments that we express with language.

Here, then, are two pieces of speculation about our species. First, we are adapted to specific kinds of emotional reactions to social situations. These reactions include moral emotions of resentment or outrage and of guilt, guided by judgments of fairness. Emotions bring characteristic tendencies to action, so that resentment, for instance, tends toward punitive action. Emotions thus affect reproduction through the actions they prompt, and so natural selection will shape the psychic mechanisms of emotion. Human emotional proclivities evolved the way they did because of this. With humans also, though, I speculate, there evolved a kind of

[6] My talk of "recipes" is drawn from Gary Marcus, *The Birth of the Mind,* a treatment of how genetic recipes lead to phenotypes.

language-infused governance of emotions. We discuss together and have linguistically encoded thoughts that work to control our feelings. On feeling a flash of resentment that you took a piece of cake that I had hoped to have, I can reason that you were as much within your rights to take it as I would have been, and so there is no cause for resentment. At this thought, my resentment may subside. If it doesn't and I complain, expressing my resentment, the rest of you may set me straight. If my resentment doesn't subside, the actions it prompts may, in my social circumstances, work in the long run to hurt my reproductive prospects. Hence come selection pressures for a genetic propensity to control emotions in certain sorts of social circumstances.

My resentment is unwarranted, I judge when you finish the cake. How does a concept like warrant work? For purposes of delineating how reasoning with such concepts can go, I suggest we think of judgments of warrant as something like plans. I plan, as it were, under what circumstances to resent people for things they do. This talk of plans for feelings sounds artificial, I admit, but when we judge that resentment would be unwarranted in my situation, the judgment acts much as would a plan, for my situation, not to resent you. Literal plans are carried out by choice, to be sure, and we can't choose what to feel. Feelings, though, do respond somewhat to judgments of warrant, as they might in the example. It's thus somewhat as if we planned what to feel, even though choice doesn't figure in the guidance of emotion in the way that plans for action get realized by guiding choice.[7]

Questions of moral right and wrong, on this picture, will be questions of what to do, but with a particular kind of emotional flavor. What is it to think an act morally wrong, as opposed just to silly or imprudent? Roughly, I propose, it is to think that the act warrants resentment on the part of others, and guilt on the part of the person who did it. Specifically moral questions, if this is right, are questions of what moral sentiments to have toward things. At their narrowest, they are questions of what to resent people for doing and what to feel guilty for doing. To guilt and resentment here, as Howard Nye has urged on me, we need to add a prospective feeling of guilt-tinged aversion toward acts we might contemplate doing.[8] This emotion is negatively valenced toward the act, and so to plan guilt-tinged aversion toward an act is to plan to be against one's doing it, in a way that has a particular emotional flavor. (Whether planning this

[7] In "Reply to Critics," I address objections to this talk of "plans" as part of a symposium with Simon Blackburn and Neil Sinclair, Michael Bratman, Jamie Dreier, and T. M. Scanlon.

[8] Personal communications.

aversion must always go with planning, all things considered, not to do the act is an important question that I won't try to answer here.)

I am contrasting, then, oughts in general and moral oughts. Narrowly moral questions of right and wrong I'm treating as at base questions of what moral sentiments we ought to have and act from. Questions in the broader class of oughts in general we call *normative* questions. These include questions of what a person ought to do all things considered. They include epistemological questions of what we ought to believe. And they include questions of how we ought to feel about things. These, I am saying, are all, in a broad, extended sense, planning questions; they are questions of what to do, to think, and to feel. Moral questions are planning questions of a particular kind, questions of how to feel about things, where the feelings in question are the moral sentiments.

Explaining Oughts

A dictum that we draw from Hume is that you can't derive an *ought* purely from an *is,* and G. E. Moore argued that oughts don't form a part of the natural world that empirical science can study. The picture I have sketched has the upshot that Moore was right. The scientific picture tells us why organisms like us would have questions whose answers can't be made a part of science. The point is that not only do we think about how things are, but we also act and feel. Our actions and feelings figure in a biological account, along with the goings-on in the head that lead to actions and to feelings about things. Questions of what to do and why, and questions of how to feel about things and why, won't figure in the picture. Yet the picture shows us addressing those questions.

Suppose I settle on helping a man in need even though I won't get any advantage from it. I think I ought to help him, and that it would be wrong not to do so, and so I help him. My coming to these conclusions must be part of any full and adequate naturalistic, biological story of me. The story, though, won't contain any fact that I've got my conclusions right or not. It doesn't contain a fact that I ought to help or that it's okay not to. It doesn't contain a fact that it would be wrong not to help or that it wouldn't be. Questions of what I ought to do and what it would be wrong to do or not to do aren't questions amenable to science. They are, I have been saying, questions of whether to help, and of how to feel about not helping. A scientific picture, then, has us asking questions that don't have scientific answers. The picture shows too why these questions aren't luxuries, but must be central questions for us. And from the scientific picture comes an account of what these questions are: they are questions of what to do and how to feel about things people do or might do. If these are the questions, we don't need to worry that they concern queer goings-on that form no part of the fabric of the universe, as John Mackie puts it.[9] They are intelligible questions, and they are questions of first importance.

I have been contrasting questions of empirical science and questions of what to do and how to feel. I should note, though, that this may not get matters quite right. Perhaps the two-way split I have made really ought to be a three-way split. First, as I've been saying, there's the empirical picture of us as special parts of the natural world, shaped as a species, as it were, by natural selection, and shaped as individuals in society by complex social dynamics, a complex human ecology. The empirical sciences of psychology, sociology, anthropology, and the like all contribute to this. Next, though, there's a part I haven't singled out: interpretation. We understand some of these natural goings-on as beliefs, assertions, plans, and the like with which we can agree or disagree. The ought part then comes third in the list, as we seek answers to the questions we can be interpreted as asking. So we have three areas of inquiry: psychosocial science, interpretation, and normative inquiry. When I speak of a person as thinking that she ought to help, and when I say that this amounts to deciding to help, I'm interpreting certain goings-on in her as the having of these thoughts.

As a first approximation, then, I'm saying, *ought* thoughts are like plans. Thinking what I ought to do amounts to thinking what to do. But this dictum needs refining. Thinking what to do can go in two stages: In the first stage, I form my valences or preferences. In the second stage, if there's more than one thing I equally and most prefer from among my alternatives, I pick one—not out of preference but out of the necessity to choose if I'm not to be like Buridan's ass. My strictly normative thinking is a matter of the first stage. We could call this part concluding what's "okay" to do and what isn't. When it's okay to do something and not okay not to, then I *ought* to do it. Thinking what I ought to do, then, is not all of thinking what to do. Rather, it's the part that matters, the valenced stage.

This ties in with a worry about the right direction of explanation. It may well be objected that I have the proper direction of explanation reversed. I started out explaining *ought* beliefs as plans. But this, even if it is right, doesn't explain normative belief in general. It doesn't explain belief in ties for what it would be best to do, the belief that more than one

[9] See Mackie, *Ethics: Inventing Right and Wrong.*

alternative would be okay. The belief that something is rationally okay to do, then, has to be explained in some other way—and once we have this explanation, it's easy to explain the concept *ought.* That a person *ought* to do a thing just means that it's okay to do it and not okay not to do it. Since we can't explain an *okay* in terms of plans, perhaps we are forced to become normative realists. We start by establishing that being okay to do is a property we can know some acts to have, and then go on from there to explain the concept *ought* and what plans consist in. That is the objection: I have tried to explain the concept *ought* in terms of plans, but the explanation, it turns out, can run only in the other direction. I answer that we can explain both concepts, *okay* and *ought,* in terms of something we do on the way to planning: forming valences. The explanation is oblique: I don't offer straight definitions of the terms "okay" and "ought" in terms of planning. Rather, I say what *believing* an act to be okay consists in. To believe it okay is to rule out preferring any alternative. It is thus to rule out a kind of valence. Normative judgments, we can say, consist in valences and restrictions on valences.

This, I'm claiming, explains the philosophically puzzling notions of what one *ought* to do and what it's *okay* to do. It explains the "to be doneness" that John Mackie thought to be no part of the fabric of the universe. It explains how G. E. Moore and other nonnaturalists could argue so convincingly that ethical thought deals with nonnatural properties. Many philosophers think that the right direction of explanation is the opposite. An answer to the question of how to live, they would say, just is a belief as to what we ought to do and what it's at least okay to do. Now of course anyone who says this has the burden of explaining what "ought" and "okay" mean. If they can't, or if their answer involves strange and incredible things like nonnatural properties, I then say that my direction of explanation is better. I start my explanation with something intelligible, with decision and the valences and restrictions that get a person to the final stage where, if need be, he goes from indifference to picking something.

Intuitions

Return now to the subject I started out with, moral intuition. I am treating moral inquiry as inquiry into how to live and how to feel, how to engage people and their actions emotionally. Often, though, moral inquiry is conducted by consulting moral "intuitions"—and indeed Henry Sidgwick, W. D. Ross, and others have argued that moral reasoning couldn't get off the ground without moral intuitions. This alleged power of moral intuition Mackie attacked as incredible, as a purported mode of knowledge that is unlike any other we know.[10] How could we be in any position to intuit moral truths, or normative truths in general? No answer is apparent in the biological picture I sketched. Nonnatural facts are absent from the picture, and so are any powers to get at nonnatural truths by intuition. Interpreting the natural goings-on as thoughts and judgments doesn't change this. If moral knowledge must depend on intuition, we seem driven to moral skepticism.

Intuitions would give knowledge, perhaps, if we had a kind of inner eye that peers into the nonnatural layout of moral facts—but that's not a picture to take seriously. Another stance we can take toward intuition is not to worry: we rely on intuition, after all, for mathematical knowledge, and so why should morality be more constrained in the ways we can know it? Now, the question of how we have mathematical knowledge is difficult. Still, at least for arithmetic and geometry, mathematics is part and parcel of empirical knowledge, the knowledge we get by counting, measuring, and the like. Our abilities to get numbers right are aspects of our abilities to get right such empirical matters as the number of pebbles in a basket. If our abilities to get morality right were like this, there wouldn't be the same puzzle about them. There would be difficult philosophical explaining to do, as with our knowledge of arithmetic and geometry, but there would be no sheer mystery as to why evolved beings like us would have powers of veridical insight in the realm of morality.

Another possibility would be that intuitions matter because the moral question just is what our moral convictions would be in reflective equilibrium, when we had given adequate heed to everything that would affect our considered moral beliefs. Moral intuitions would matter, then, as the starting points for reaching reflective equilibrium. I'm claiming, though, that moral claims aren't claims in interpreted psychology. The question of what we would think if such-and-such conditions obtained is mostly an empirical one, along with the further question of how to interpret the state we would then be in. I have been saying that the moral question isn't what we *would* think in such-and-such conditions, but what to do and how to feel about things we do or might do. These questions aren't answered by interpreted psychology alone.

[10] Ibid. On the necessity for intuitions, see Sidgwick, *The Methods of Ethics;* and Ross, *The Right and the Good,* esp. 39–41.

Now it might seem that I have escaped the problem of relying on intuitions. If normative thoughts are plans, or valenced restrictions on plans, then to come to normative conclusions, we just have to plan. This, however, doesn't free us from intuitions. As we plan, we'll weigh considerations for and against actions. Take even a case of nonmoral planning, thinking whether to go to the store. In favor I might weigh the consideration that there I can get cigarettes. I can go on to evaluate whether to weigh that consideration in favor. I settle what to weigh into my decision and how, and form a string of considerations that support other considerations. At some point, though, the string comes to an end. Perhaps I weigh the fact that I'd enjoy smoking in favor of smoking, on no further ground. And perhaps I weigh the chance that I'd suffer if I smoked against a plan to smoke. Weighing enjoyment in favor and suffering against, on no further ground, amounts to having an intuition about why to do things. Intuitions, then, apply to planning, and not just to thinking how things stand. If I keep challenging my thoughts about what to do and why, I end up grounding my planning in intuition.

I accept, then, that normative thinking rests on intuition. This seems to raise the same question again: why think we can intuit why to do things? Like questions go for thinking how to feel and why: why think we can intuit why and why not to feel certain ways about things? But thinking of ought judgments as plans leads to an answer. I intuit, we said, that the chance that I'd suffer if I did a thing is reason not to do it. But to say that I have this intuition is just another way of saying that I confidently weigh the chance of suffering against doing a thing, and on no further ground even if I ask myself why.

To say this is to use the term "intuition" in an empirical, nonnormative sense, as Haidt does—as a certain kind of state of mind that is open to empirical study. We could instead use the term, though, in a normative sense: an intuition, we could say, is a state of mind of accepting something, not on the basis of further reasoning even upon challenge, that we ought to place some trust in. To think something an intuition in this sense is to plan to rely on it. I'll call intuitions in the nonnormative sense in which they figure in psychology "de facto" intuitions. These are judgments made confidently, on no further grounds, with no felt need for further grounds even upon challenge. Intuitions in the normative sense I'll call intuitions "de jure." These are de facto intuitions to rely on. It's a normative claim, then, that de facto intuitions are genuine intuitions—and a claim that we need, I have been saying, for coherent planning.

Ideal Conditions

I have been stressing the distinction between nonnormative psychological questions of how we do form moral judgments and normative questions of how we ought to. What we will plan under what conditions is a psychological question, whereas normative questions are planning questions of what to do. The two are closely tied to each other, though. We can ask the planning question of when to trust our own planning. We can ask under what conditions to trust our planning most. That amounts to asking what conditions are ideal for planning. Ideal conditions, we might conclude, involve such things as full information vividly taken in and contemplated, and an alert, engaged, and dispassionate frame of mind. If we come to a precise view about what those conditions are, we can then ask the psychological question of what, in those conditions, we would plan.

I face a moral dilemma, suppose—I'll give as an example a simple and far-fetched dilemma that I'll talk more about in the next lecture. A father stands on the bank of a river where two canoes have capsized with children in them. His own daughter was in the far canoe, and he can rescue her. Alternatively, though, he could rescue two children of strangers who are nearer to him. He can't do both; what ought he to do?

This first is an ought question; now we can ask another kind of question: how would we answer this first question in ideal conditions for judgment? If we get an answer to the second question, which is psychological, we'll come to an answer to the first. Suppose I conclude, "Under ideal conditions for judgment, I'd judge that he ought to rescue his daughter, even though that means rescuing only one child when he could have rescued two." Relying on myself as I'd judge in ideal conditions, I can now say, "He ought to rescue his daughter instead of the other two children."

It's not that the moral conclusion is *entailed* by a finding in interpreted psychology. Rather, what's going on is this: When we call conditions for judgment "ideal," we mean that judgments in those conditions are ones to trust. To accept this is to plan to trust such judgments. So I accept the claim, imagine, "In ideal conditions, I would judge that the man ought to rescue his daughter." Equivalently, I accept this: "The judgment that he ought to rescue his daughter is one to trust."

To accept this is to plan to trust this judgment, the judgment that the man ought to rescue his daughter. To trust the judgment means being disposed to emulate it in one's own judgment. So following through on the plan, I make the judgment "The man ought to rescue his daughter."

If, then, we could settle under what conditions to trust our normative judgments, then we could come to normative conclusions on the basis of interpreted empirical findings. From the empirical finding that in those conditions for contemplation I'd judge the man ought to rescue his daughter, I could reason to judging that he ought to rescue his daughter, and voice my state of mind by saying that he ought to rescue his daughter. This isn't deriving an *ought* from a psychological *is* alone, for there's an intervening normative premise. The premise amounts to this: that what I'd *find* wrong in those particular conditions *is* wrong—that what I would then *think* ought to be done *ought* to be done.

Possibly, then, we could find a systematic way to move from psychological findings to moral conclusions. In many kinds of cases, after all, a transition from *is* to *ought* is entirely obvious and uncontroversial. If you're driving late to work and a child will be killed unless you stop, then you ought to stop. How to weigh a child's life against arriving promptly at work is something we've settled beyond need for further review. If the conditions under which to trust one's normative judgments were similarly unproblematic, then the problematic parts of ethics would be reduced to questions of interpreted psychology. The move from *is* to *ought* still wouldn't be one of entailment, but it might be systematic and trustworthy. We aren't at that point yet, though—and if we did get there, it would still be important to distinguish *ought* questions from psychological questions, to keep track of what we had achieved and what our basis was for accepting the ought conclusions we accepted.

Coherence and Inconsistency

Plans, I claimed, require intuitions, but I need to make this claim more precisely. At a moment, I can find it clear that the fact that I'd enjoy something weighs in favor of doing it. I can then rely on this as a premise without relying on the further psychological premise that I find this obvious. No thoughts about intuition enter into my thinking, and I haven't skipped over any steps that would be needed to make my thinking explicit and fully cogent. Over time, though, I can plan what to do only if, at least implicitly, I do place some stock in my earlier conclusions without rethinking them. I trust my earlier conclusions, and I can't be justified in doing this unless the fact that I earlier found something obvious on no further ground is at least some reason to accept it. Planning requires thinking that the *is* of interpreted psychology—that I implicitly accept an ought, and would accept it explicitly if challenged, on no further ground—supports accepting the *ought.* Not only must I have de facto intuitions, but I must also trust them; I must treat them as intuitions de jure.

I don't mean, though, that de facto intuitions are to be trusted entirely. Seeming intuitions can clash, and indeed seeming intuitions about what to do can clash severely. The trust to put in them can only be defeasible. Even if moral claims didn't mean what I say they do, and even if the visual model held good for intuitions of moral right and wrong, we'd have to test intuitions against each other and revise them in light of conflicts. Philosophical work on normative ethics, much of it, consists in engaging in this refinement of intuitions—but there's no end of controversy as to where the weight of corrected intuition falls.

I have been suggesting that we might get further by conceiving of our questions as ones of what to do and how to feel about things, and why. This won't end our dependence on intuitions, but we can see if the intuitions we now rely on are more tractable. Much of what I'll be doing in the next lecture will go over ground familiar to moral philosophers, and we'll have to hope that the resulting treatment makes contact with ordinary moral thought, or there would be little reason to trust it. A lot of what I'll be saying in the next two lectures stems from decision theory and from arguments that decision theorists have made. We can think of decision theory as a systematic development of intuitions about what to do and why.

Decision theorists in the classical Bayesian tradition work to formulate what it means to be consistent in one's policies for action, and then derive surprisingly strong results from the conditions they lay down. This tradition stems from the work of, among others, L. J. Savage, who rediscovered a way of thinking that had been developed by F. P. Ramsey toward the end of his short life.[11] If a way of making choices satisfies the Savage conditions (or conditions in a like vein), as it turns out, then it is as if one were maximizing an expectation of value. It is as if, that is to say, one had numerical degrees of credence and numerical evaluations of the possible outcomes, and acted to maximize expected value as reckoned in terms of these evaluations and degrees of credence. (The term "expected value" doesn't here mean what it would mean in ordinary English; one acts as if to maximize

[11] Classic developments of decision-theoretic arguments are Ramsey, "Truth and Probability"; and Leonard J. Savage, *The Foundations of Statistics* (1957). Peter Hammond, in "Consequentialist Foundations for Expected Utility," develops a framework in terms of sequential decisions, and this offers, I think, the clearest case that departing from the strictures of classical decision theory is incoherent. Unfortunately, Hammond's argument is couched in fearsome mathematical apparatus.

an expectation in the mathematical sense, summing up one's evaluations of the possible outcomes each weighted by one's degree of credence that it would be the outcome.) Bentham the hedonist was right at least formally, it seems to follow: if one's policies for action are consistent, one acts, in the face of uncertainty, to advance the good on some scale of evaluation. The scale may not gauge pleasure, but there will be some such scale or other.

The conditions that classical decision theorists put forth as innocuous and compelling, though, combine in ways that clash with strong intuitions. They are for that reason controversial; critics look to show that not all the classical conditions are genuinely demands of reason. In the lectures to come I rely heavily on the findings of classical decision theory, and so although I won't scrutinize the controversies in any depth, I'll glance at one famous example, due to Richard Zeckhauser.[12]

You are forced to play Russian roulette, but you can buy your way out. What is the most you would be willing to pay, the question is, to remove the bullet, reducing your chance of shooting yourself from one in six to zero? Or that's the first question; once you answer it, we ask a more complex one. You are instead, it turns out, forced to play a worse version of Russian roulette, with four bullets in the six chambers. What's the most you would pay, the question now is, to remove *one* of the four bullets? In particular, is it more or less than before?

Most people answer less. But you should pay *more,* goes an argument from orthodox decision theory. This problem is equivalent, after all, to a two-stage problem, as follows: In the first stage, you are forced to play with three bullets and no chance to buy yourself out. In the second stage, if you survive, you are forced to play with two bullets, but you can pay to remove both. The amount to pay in the second case, then, is anything you would pay to remove both of two bullets if they were the only two bullets —surely more than to remove one sole bullet.

This case and others like it have been staples of debate on the foundations of decision theory, and ways out of this conflict of intuitions have been proposed. The first thing to note, though, is that the intuitions in conflict are strong. Is removing two bullets worth more than removing one, if in each case you thereby empty the pistol? Surely. Does anything matter, in these choices, but chance of surviving and how poor you will be if you do? Not much; those seem the predominant considerations. It

[12] The example is presented in Kahneman and Tversky, "Prospect Theory: An Analysis of Decision under Risk," 283. It is a version of the famous "Allais paradox" for classical decision theory.

doesn't matter, then, whether you must play the four-bullet game or the two-stage game, since they involve choice among the same chances of death. Does it matter if you choose at the start of the two-stage game what to pay if you survive the first stage, or decide once it turns out you have survived the first stage? Clearly not. Orthodox decision theory goes against intuition for this case, but any alternative to orthodoxy will violate one of the strong intuitions I just voiced. The constraints in classical decision theory that do the real work are all exemplified in the argument I just gave, and so at least for cases like this one, if the argument is good, then classical decision theory is pretty well vindicated.

I myself am convinced that what we gain in intuitiveness when we depart from the orthodox views in decision theory in cases like this is less than what we lose. That would be a long argument, though, and I can't expect you to accept this conclusion on my say-so. What I hope you are convinced of is that some of our strong intuitions will have to go whatever we hypothetically decide to do in the Zeckhauser case. In the lectures that follow, I'll proceed as if the conclusions of orthodox decision theory are right—but you should note that part of the argument remains to be discharged, and it is controversial among the experts whether it can be.[13]

I'll be assuming without further argument, then, that the constraints of decision theory are ones of consistency in action, or something close to it. Whether they are full-fledged matters of consistency is a tricky question, and so I'll use the word *coherence*. Why, though, does coherence in plans for action matter—especially when they are plans for wild contingencies that we will never face, like being forced to play any of several versions of Russian roulette? With questions of fact, the problem with inconsistency is that when a set of beliefs is inconsistent, at least one of the beliefs is false. I'm suggesting that we think of ought questions, in the first instance, as planning questions. Answers to them may in the end count as true or false, but we don't start our treatment with talk of truth and falsehood and help ourselves to these notions in our initial theorizing. With incoherent plans, I accept, the oughts we accept in having those plans can't all be true, but that isn't at the root of what's wrong. So, indeed, what *is* wrong with incoherent plans?

As a first approximation, I can say, incoherent plans can't all be carried out. If I plan to be here today and also plan to be on top of Mount Kenya,

[13] For critiques of classical decision theory with references, see, for instance, Amartya K. Sen, "Rationality and Uncertainty"; and Edward F. McClennen, *Rationality and Dynamic Choice.*

believing that I can't be both places on the same day, my beliefs and plans are inconsistent. Either, then, my belief that I can't be in both places is false, or one of my plans I won't carry out no matter what choices I make. Some of the plans I'll be talking about in the next lecture, though, are wild contingency plans that I'll never be in a position to carry out anyway. I might talk about such wild plans as a plan for what to prefer for the contingency of being Brutus on the Ides of March. And some of the states of mind that can be coherent or not with others won't be simple plans but constraints on plans and beliefs—that, for instance, I plan to pay more, if forced to play Russian roulette, to empty the pistol of two bullets than of one.

The problem with inconsistent plans is that there is no way they can be realized in a complete contingency plan for living. For each full contingency plan one might have, something in the set will rule it out. Or more completely, we'd have to talk about inconsistent beliefs, plans, and constraints. If a set of these is inconsistent, there's no combination of a full contingency plan for living and a full way that world might be that fits. And judgments get their content from what they are consistent with and what not.

Preview

In this first lecture I have contrasted biological thinking about us and the normative thinking that the biological picture has us engaging in. A rich enough biological picture, I think, explains why a highly social, linguistic species like ours would engage in normative thinking and discussion, and in moral thinking and discussion in particular. I also talked about intuitions. We couldn't coherently proceed with normative thinking without giving some trust to some of our de facto intuitions, treating them as intuitions de jure. (Indeed, I would claim that this applies to thinking of all kinds—but I haven't gone into that in this lecture.) At the same time, some of our strong intuitions are inconsistent with each other, and so our trust in de facto intuitions, to be coherent, must be guarded.

In lectures that follow, I'll take this very high-level normative thinking about intuitions and reasoning, and turn to morality. Our lives are social, and a large part of thinking what to do and how to feel is thinking how to live with other people. We address these questions partly each by ourselves and partly together in discussion. I'll be keeping my eye on moral thinking as thinking how to live with each other, and on the question of how to regard our moral intuitions. The moral argument that I pursue and scrutinize is one that may be very powerful, but that raises difficult questions.

This is an argument that owes the most to Berkeley's late John Harsanyi. It leads to conclusions that clash with strong moral intuitions, and I'll be trying to think through the force of these intuitions. In the two lectures that follow, then, instead of just describing moral thinking as thinking how to live with each other, I'll engage in moral thinking in a reflective and highly theoretical way.

II. LIVING TOGETHER: ECONOMIC AND MORAL ARGUMENT

We are beings who think about how to live. We live each with others, and we think how to live with each other. Sometimes a person will think about such things by herself, and sometimes we think and discuss together. These are truisms, but I argued in the first lecture that the truisms are rich in consequences. They explain, if I am right, the philosophically puzzling area of thought we call "normative," thought that somehow involves oughts.

I want to ask in this lecture and the next whether such a self-understanding could have any bearing on questions of right and wrong, of good and bad. In the first lecture I talked about moral concepts without using them. I did metaethics, not normative ethics, not the work of thinking through what *is* right and what is wrong, and why. My metaethics leaves room for any coherent answer whatever to normative questions of what's right and what's wrong to do—and a wide range of possible answers are coherent. I want, though, to explore whether the metaethical picture I sketched contributes at all to making some answers to normative questions more plausible than others. In doing so, I'll have to pass lightly over controversies familiar in the literature of ethical theory, giving quick and insufficient arguments on issues that have been extensively and subtly debated.

A Social Contract and the Strains of Commitment

My late colleague William Frankena finished his short book *Ethics* with the dictum "Morality is made for man, not man for morality."[1] His saying is widely quoted. He told me that he regretted ever saying this, but I don't see that he had anything to regret. If morality should matter to us, if we should adhere to moral demands even at great sacrifice, then morality shouldn't be arbitrary. Concern for morality should be out of concern for

[1] Frankena, *Ethics,* 98.

something that makes morality of value—and how could that thing be anything other than being of value for people? (I don't mean to rule out other sentient beings, but in these lectures I'll stick to people.)

Most philosophers, I think, will agree with Frankena's saying, but we fall into contention when it comes to drawing out its implications. Moral inquiry in philosophy often comes in either of two broad styles. One is humanistic and pragmatic, thinking what's in morality for us, for us human beings, and asking what version of morality best serves us. The other broad style is intuitionist, in one important sense of that term: consult our moral intuitions, revise them as need be to achieve consistency, and embrace what emerges. The point isn't that these two styles of moral inquiry need entirely be at odds with each other. The hope in consulting and systematizing intuitions is that doing so will uncover a deep, implicit rationale for our intuitive responses, and that the rationale we discover will turn out to be a worthy one. The hope is thus that, carried out in the right way, the two broad styles converge. Humanistic pragmatists start out with a vague rationale for ethics, a value ethics has that can be appreciated in nonethical terms. As Henry Sidgwick argued more than a century ago, however, a morality made for humanity must in the end be grounded on some intuition—an intuition, perhaps, as to how humanity matters.[2] His vision was, then, that the two approaches, pragmatic and intuitive, amount to the same approach. Still, initially at least, the two are quite different in spirit.

If morality is for humanity, then we might expect utilitarianism to be right. Moral rules, we might expect, will tell us each to act for the benefit of all humanity. The right act will be the one with the greatest total benefit to people. Utilitarianism, though, notoriously conflicts with strong moral intuitions. As a simple example, I'll start with the case in the first lecture of children drowning. I'll then broach a line of argument that appeals to other intuitions and seems to lead back to the utilitarian answer. The case illustrates a much broader, systematic argument for utilitarianism, one that draws on decision theory and was most notably advanced by Berkeley's own John Harsanyi well before he came to Berkeley. Aspects of the argument have been widely debated, and my aim is to use the debate to explore how moral inquiry might proceed if it consists in doing the sort of thing I claim, in thinking how to live together.

The case is due to Diane Jeske and Richard Fumerton.[3] Two canoes of children capsize in rapids, and a man on the bank can rescue some but

[2] See Sidgwick, *The Methods of Ethics.*

[3] Jeske and Fumerton, "Relatives and Relativism."

not all of the children. Close to him are two children, and he could rescue both. Farther away is his own daughter, and alternatively he could rescue her but not the other two. Utilitarianism seems to say that, faced with this grim choice, he should rescue the two children rather than the one. Many people have the strong intuition that the father is morally permitted—perhaps even required by the duties of parenthood—to rescue his daughter, even though he must then let two children drown instead of one.

This example is contrived, in the style of many philosophical examples. The hope is, though, that such examples can let us examine considerations in isolation that get too complex to handle clearly in the kinds of morally fraught situations we are most apt to encounter.

I'll now introduce the style of argument that I'll be exploring. Imagine now that the situation is a little more complex. There are two fathers on the two riverbanks by the rapids, and two canoes are swamped, each with two children. For each father, his own children are in the farther canoe. Each could save either the two children closest to him or one but not both of his own children in the far canoe. The rule to give preference to one's own children, if each father follows it, means that each father loses a child. The rule to save as many children as possible, regardless of whose they are, means, if each father follows it, that no father loses a child.

Perhaps in this freak circumstance, the two fathers could quickly reach an agreement that each would rescue the other's children. They would then each have a contractual obligation to act as utilitarianism would command, and for this case, the contrast between intuition and utilitarianism might disappear. In its prescriptions for this particular case, a social contract would thus coincide with utilitarianism.

Return, though, to the first story, with one father whose child was in the far swamped canoe. Suppose that in advance, the two fathers contemplate this contingency. One of them will be on the bank, with one of his two children swamped in the far canoe. Both children of the other will be in the near canoe. The man on the bank will be able, then, to save either both of the other father's children or one of his own. The fathers might come to a social contract covering this eventuality. What would it be? Suppose first that they agree that each is to save his own in preference to saving both the nearer children. If they know the agreement will be kept, then each stands to lose both of his children if he's the unlucky father who has two children at risk and can't rescue either one, and to lose no child in case he's there to do the rescuing. Next, suppose instead they agree that each will rescue as many children as he can. Then if the agreement will be kept, each stands to lose one child if he's the unlucky father on the bank, acting on his agreement, and to lose no child if he's the lucky absent father whose children get rescued by the other. In short, then, so long as whatever agreement they reach they will keep, then the first agreement in the unlucky case means losing both one's children, whereas the second in the unlucky case means losing only one child. Each is a terrible loss, but losing both children is even worse that losing one—and the two cases are equally likely, we have supposed. For each father, then, the second agreement offers the better prospect.

Again, then, for the case in question, two approaches give the same answer. Utilitarianism says to rescue as many children as one can, and so does the social contract that people would make if they knew that the social contract would be kept.

This kind of argument generalizes. John Harsanyi in the 1950s proved two famous theorems that apply—theorems that I think should be more famous than they are among students of moral philosophy. The import and the limitations of his theorems have been debated in the philosophical and economic literature, and I'll be exploring how some aspects of the discussion might go if moral inquiry is the sort of thing I think it is: planning how to live with each other.

First, though, let's explore further the case of the children and the swamped canoes. The two fathers have agreed what to do in the contingency, and now one of them finds himself in the dilemma on the riverbank. He has agreed to save the two children that aren't his, but still, of course, he is strongly moved to save his own child. What motive might he have to keep the agreement and let his own child drown? His motive might be one of fair reciprocity. "He would have done the same for my two children if our positions had been reversed," he can say to himself. Still, he faces the question of whether to reciprocate. Possibly, fair reciprocity will insufficiently motivate him, and he will fail to reciprocate, in this desperate situation, what the other father would have done for him and his children. A further question arises too, then: would the other father have been sufficiently motivated? If the other would have reneged had their positions been reversed, then the father on the bank loses his rationale from fair reciprocity.

Here, then, the upshot of contractarian thinking deviates from that of utilitarian thinking. Suppose for now that I am right that, if the two could make an agreement with full assurance that the agreement would be kept, they would agree on the arrangement that utilitarianism prescribes. In this way, utilitarianism can stem not only from motives of benevolence but also from motives of fair reciprocity. That's only, though, if the motivations of fair reciprocity are never overwhelmed by other motives and the parties have full assurance of this.

A contractarianism that heeds the limits of motives of fair reciprocity will be quite different. What would we have agreed on, under the constraint that the motivations we would have to keep the agreement would be sufficiently strong, if everyone knew that the agreement would be kept? That will depend on a psychological question: how strong are motives of fair reciprocity? How strong can we trust them to be under various relevant conditions?

We can see now why there might be a contractarian excuse for rescuing one's own child in preference to two others. If we had been able to make any agreement whatsoever and make it effective, we would have agreed to rescue as many children as possible, no matter whose. But we can't produce such strong motives—and under the constraints of how strong motives of fair reciprocity can be, we wouldn't have made such an agreement only to expect it not to be kept.

I'm touching on what John Rawls called the "strains of commitment."[4] In most of the rest of this lecture, I'll ignore them and explore other questions. I'll consider contractarian arguments that assume full assurance of full compliance, severe though this limitation is. Any full exploration of contractarian arguments, utilitarianism, and moral intuitions, though, would have to pay great heed to the strains of commitment.

The Separateness of Persons

One way of arriving at utilitarianism is to say that morality consists in benevolence, in impartial concern for all involved, including oneself. Rawls responded that impartial benevolence is a weak motive, and that a far stronger motive is fair reciprocity. T. M. Scanlon puts the motive differently: roughly, as a concern to live with others on a basis that no one could reasonably reject.[5] The canoe case suggested a way in which all these might coincide, at least in the case of full compliance. Fair reciprocity consists in abiding by a practice if it's the practice we would have agreed

[4] Rawls, *A Theory of Justice,* 145, 177–78, 490–504. My discussions of Rawls in these lectures refer to this version of his theory, although I think they apply to later versions as well.

[5] Ibid., 494–95, 500; Scanlon, *What We Owe to Each Other.* More precisely, contractualism, Scanlon tells us, "holds that an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement" (153).

to before we knew who would be in what position. To such a practice, no one could reasonably object.

The question we ask in moral inquiry, I have been saying, isn't the psychological one of what motives we *do* have and how strongly, but the question of what motives *to* have. It's a planning question, a question of how to live with each other. Nothing in the metaethics that I have laid out dictates an answer. Still, the ideals of fair reciprocity and of living with others on a basis they could not reasonably reject seem good candidates for what to want in one's dealings with others. These aims are vague, but I propose to think together with people who might be brought to share these aims, and try to work toward specifying them in a way that might make them worthy of pursuit.

Morality, it is often said, is grounded in respect for persons, and utilitarianism fails in that it can prescribe actions that violate people's rights and fail to respect them. I can't, of course, go over the history of systematic attempts to ground morality in respect and get nonutilitarian conclusions, but my own reading of the history is that these attempts have not had great success—and our brief discussion of the canoe case illustrates why coherent, nonutilitarian theories are so elusive.[6] The vague aims of fair reciprocity and of dealing with others in a way that no one could reasonably reject do strike me as good places to start in working out what aims to have, and what we would have agreed on seems highly relevant to respect and what it demands. I'll be arguing in these lectures that these starting points lead to a moral view that is utilitarian in form, but as I say, considerations of respect are widely thought to tell against utilitarianism. Before I scrutinize contractarian arguments further, I'll say a few things about why I don't think respect leads us straightforwardly in directions that oppose utilitarianism.[7]

Utilitarianism, it is sometimes said, ignores the "separateness of persons."[8] One person's gain doesn't compensate for another's loss. A person is not to be sacrificed for the sake of another. Thinking in terms of "gains"

[6] Kant, *Grundlegung der Metaphysic der Sitten.* Holly M. Smith, "Rawls and Utilitarianism," is a fine discussion of whether Rawls succeeds in finding a rationale for a nonutilitarian theory. In my "Morality as Consistency in Living: Korsgaard's Kantian Lectures," I look at Korsgaard's attempt to derive a Kantian morality; see Christine M. Korsgaard, *The Sources of Normativity.*

[7] W. K. Frankena, in "The Ethics of Respect for Persons," argues that the content of morality must be settled in order to determine what constitutes treating a person with respect, so that the demands of respect can't be the basis of morality.

[8] See Rawls, *A Theory of Justice,* 26–31. "The most natural way" to arrive at utilitarianism, Rawls says, is "to adopt for society as a whole the principle of rational choice for one man." This is not the only way, he says, but he concludes, "Utilitarianism does not take seriously the distinction between persons" (26–27).

and "losses" or of "sacrifice," though, requires a base point of comparison, and so we'll need some rationale for heeding one possible base point as opposed to others. Suppose we have persons Ida and Jay and states *A* and *B,* with Ida better off in state *A* and Jay better off in state *B.* Let's give numbers to how well off they are:

State	*A*	*B*
Ida	9	5
Jay	1	3

Ida's gain in going from state *B* to state *A* doesn't compensate for Jay's loss, we might try saying: Ida gains, going from 5 to 9 for a gain of 3, but Jay loses, falling from 3 to 1. Jay has only one life to lead, and we can't sacrifice him for Ida's benefit. If we frame matters differently, however, we come to the opposite conclusion: In going from state *A* to state *B,* Ida loses. Jay gains, to be sure, but he and Ida are separate persons, and Ida can't be sacrificed for Jay.

To choose between these two seeming upshots of the separateness of persons, we must choose between state *A* and state *B* as the base state from which "gains" and "losses" are measured and "sacrifice" is attributed. Rawls seemed to choose the state with the worst-off person—state *A* in this case. That might raise the worry of whether we can legitimately "sacrifice" the well off to benefit the badly off. Robert Nozick and some others who appeal to Kant say that we choose as the base state for comparison the state in which people are entitled to what they would have.[9] Rawls replies that when the basic structure of society is at issue, we're asking what entitlements to institute.[10] Intuitions that invoke ownership and other entitlements are very strong, and they may well be "wired in" to the human psychic makeup.[11] They are very sensitive, though, to "framing" effects: even a person's self-regarding choices are affected by attributing

[9] See Nozick, *Anarchy, State, and Utopia.* Nozick discusses whether the rich don't have the same right to complain of Rawls's difference principle as the poor would have to complain of making the best off as well off as possible (190–97). Elsewhere, he speaks of "sacrifice" (32–33).

[10] Rawls, "The Basic Structure as Subject." See also Rawls, *A Theory of Justice,* 7–10.

[11] Alan Page Fiske, in "The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations," argues that people use four "elementary relational models" to coordinate social interaction; he calls them Communal Sharing, Authority Ranking, Equality Matching, and Market Pricing. The universality suggests genetic adaptation to think of social relations in terms of these schemas. Concepts of property are obviously involved in Market Pricing, and they are also involved in aspects of some of the others. Gifts come under Communal Sharing, and Equality Matching includes matching of contributions.

ownership.[12] (Consider “endowment effects”: We “give” a person a coffee mug and ask him if he’ll trade it for a chocolate bar. He says no. It seems he prefers having the mug to having the chocolate bar. But if we had given him the chocolate bar instead, he would have refused to trade it for the mug. It seems he would then prefer having the mug to having the chocolate bar. The only difference is which of the two objects he frames as already “his.”)[13] Can we find some basis for attributing entitlements, then, that is independent of a pragmatic test, independent of evaluating the consequences of a system of entitlements, by a standard that doesn’t assume the importance of the entitlements in advance? Nozick tried, but he left the basis of what he was saying unexplained, and seemed to appeal to the pragmatic advantages of systems of property.[14]

I conclude that we can’t talk of “gains,” “losses,” and “sacrifice” until we identify some base point for the comparisons. It is true enough that we are separate persons—but nothing about what we may permissibly do follows from that by itself. Our strong intuitions do latch on to some base point or other, but not in any consistent way. Perhaps we could establish some base point as morally relevant. One way to do so, though, would be the way I’ll be exploring: ask what we would have agreed to from behind a veil of ignorance, what we would have agreed to treat as morally relevant.

Harsanyi’s Theorems

The point of morality, I’m taking it, is to live with each other on a basis that none of us could reasonably reject. No one has a reasonable objection if the system we live by is what we would have agreed to in fair conditions—and one way to make conditions fair is a veil of ignorance. We saw in the canoe case that this may yield utilitarian prescriptions. Harsanyi argued that this upshot generalizes.

His argument starts with the coherence of plans for action as elucidated by classical decision theory. As I discussed in the first lecture, decision theorists have shown that if a way of ranking actions satisfies certain conditions, then it is as if the person chose by maximizing an expected value.[15] It is as if the person formed degrees of credence in the relevant

[12] On “framing” effects, see A. Tversky and D. Kahneman, “The Framing of Decisions and the Psychology of Choice.”

[13] See Daniel Kahneman, J. L. Knetch, and R. H. Thaler, “Experimental Tests of the Endowment Effect and the Coase Theorem.”

[14] Nozick, *Anarchy, State, and Utopia.* For critiques along this line, see Hal R. Varian, “Distributive Justice, Welfare Economics, and the Theory of Fairness”; and Gibbard, “Natural Property Rights.” See also Sidgwick, *The Methods of Ethics,* 278–83.

[15] Ramsey, “Truth and Probability”; Savage, *The Foundations of Statistics.*

eventualities, attributed levels of value to the various possible outcomes, and then took the alternative that held out the greatest expectation of value, reckoned with those degrees of credence and levels of value. By the "standard conditions" I'll mean any of the various sets of conditions that have been shown to yield the result, and "coherent" plans, I'll assume, are plans that satisfy these conditions. As I indicated in the first lecture, it is highly contentious whether the axioms are requirements of coherence in any ordinary sense of the term, but I'll be exploring what we should think if they are.

Harsanyi proved two theorems that I'll call his two welfare theorems. His first welfare theorem concerned something like Rawls's "original position" with his "veil of ignorance."[16] Think of valid moral rules as the rules one would choose assuming an equal chance of being anyone. Assume one's preferences are coherent, in that they satisfy the standard conditions. Then one will prefer the rules that would yield the greatest total utility. Here by "individual utility," I mean the scale that represents one's preferences given that one will turn out to be that person.[17]

Harsanyi's second welfare theorem is this: Suppose that prospective individual benefit is coherent, and so is desirability from a moral point of view. Suppose also that morality is for humanity in at least the following sense: if one prospect is better than a second for each individual, it is the better prospect ethically. (This is a version of what is called the *prospective Pareto condition.*) Then desirability from a moral point of view, he proved, is a weighted sum of individual benefit.[18] The only way ethical evaluation could satisfy these conditions and depart from utilitarianism is by weighing one person's benefit more than another.

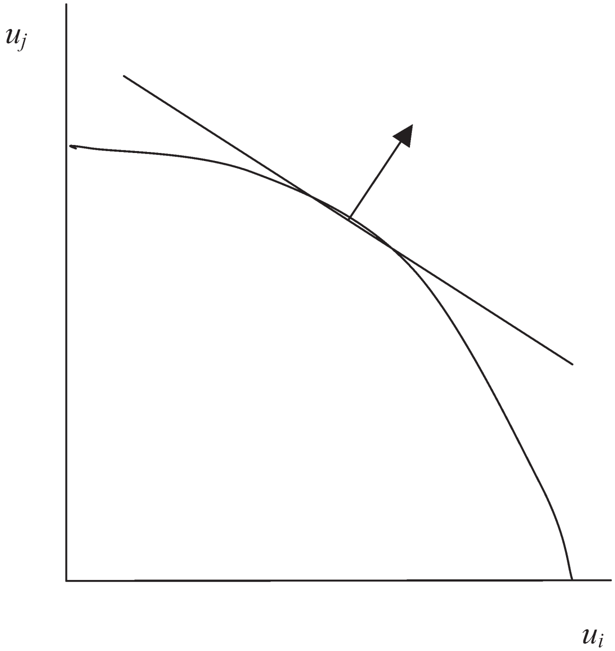
Economists represent the theorem in graphical form. We take the simple case of two people. Each social order we might have instituted gives

[16] Rawls used these terms in *A Theory of Justice.*

[17] Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." The term "utility" is often defined differently from the way I define it here; one frequent meaning would let the person's "utility" be the scale that represents the preferences of that person himself. I consider later in this lecture why the two senses might diverge.

[18] Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," 313. See John Broome, *Weighing Goods: Equality, Uncertainty and Time,* for a superb treatment of this theorem and its ethical implications. Harsanyi spoke of individual "preferences," and implicitly assumed that they characterize individual good or benefit; I discuss this shortly. Broome distinguishes questions of good and questions of preference. He confines the term "Pareto condition" to one couched in terms of preferences, and calls the principle as I have stated it the "Principle of Personal Good" (155). Harsanyi's two theorems, understood as abstract mathematical results, can of course be given more than one interpretation in their applications to ethical questions, and each interpretation would require separate scrutiny. Broome calls Harsanyi's second welfare theorem, under the interpretation I am giving it, the "Interpersonal Addition Theorem" (162–63).

each person a prospective benefit, and we can represent this benefit by a point, with Ida's benefit the *X* coordinate and Jay's the *Y* coordinate. These, we can say, are the combinations of prospects that were feasible. The feasible combinations that satisfy the prospective Pareto condition, such that no alternative would have given both people better prospects at once, lie along the frontier at the upper right. A moral theory that is consistent with the prospective Pareto condition chooses one of the points on this frontier as that of the just social order. This point, though, maximizes some weighted combination of the individuals' prospective benefits. Graphically, we can see that it is maximally extreme in some direction. Harsanyi's second welfare theorem is that a combination that satisfied his three conditions has this property.



The challenge to anyone who wants to get a nonutilitarian morality out of thought on a social contract is how to evade the force of Harsanyi's two theorems. If you are going to be a nonutilitarian, you will adopt moral rules that none of us would have chosen for his own sake unless he knew of some special way that he and not others stood to benefit. And any evaluation of the prospects that various different moral orders bring must either (1) violate some demand of rationality, (2) weigh one person's utility above another's, or (3) rank some prospect best even though another one prospectively benefits everyone more.

Now Harsanyi's two welfare theorems have been much discussed, if not sufficiently. The quick, careless statements of the theorems that I have given would require close scrutiny, and important parts of the needed scrutiny are in print.[19] What I can hope to do is just to select a few issues that are relevant to these debates, framing the theorems as parts of moral inquiry as I have been picturing it.

A Person's Good

What is a person's good or benefit? In the tradition that Harsanyi worked in, a person's good is a matter of her preferences. We gauge the strength of her preferences by the preferences she would have among risky prospects: if you would risk a one-in-a-million chance of being killed by a car, but no more, to cross the street and buy a chocolate bar, then the benefit to you of a chocolate bar counts as one-millionth the harm to you of being killed. This notion of benefit has three chief problems. One seems tractable enough: philosophers hasten to add that the preferences must be considered and informed. A second problem is more difficult: the preferences of any decent person won't just be for that person's own benefit. The person will care about others; he will care about fairness; he will care, perhaps, about living with others on a basis that no one could reasonably reject. A person's benefit is at best one component of his considered, informed preferences. What component is that? The third problem interacts with the second. What people prefer for themselves differs from person to person. Some differences aren't fundamental: I dislike asparagus and my wife loves it, and so I prefer not to eat it and my wife prefers to eat it. Basically, though, we both want to enjoy our food, and we're different in what we enjoy. For the case of being her with her tastes, I want to eat asparagus. Other examples, though, might be different. When I lived in Ghana, people told me that one thing they set great store in was a big funeral. That puzzled me, but as I thought about it, I realized that a big funeral indicates how one is loved and respected in the community, and to be loved, respected, and missed did seem to me to be things a person could intelligibly put great stock in. Still, once we distinguish carefully what goes on in one's life and what happens after, people may differ in whether they care, for their own sake, how they are regarded after death.

When I stand behind a veil of ignorance and choose a social ethos to institute, I contemplate that I may turn out to be you with your basic

[19] For critiques of Harsanyi, see especially Broome, *Weighing Goods.* See also David Gauthier, "On the Refutation of Utilitarianism"; and Kotaro Suzumura, "Interpersonal Comparisons of the Extended Sympathy Type and the Possibility of Social Choice."

preferences, and I may turn out to be me with my basic preferences. You and I may differ even in the basic preferences we would have if our preferences were fully considered and informed. Rawls stressed this, and proposed that instead of looking to self-regarding preferences, we look to what he called an "index of primary social goods." Primary goods he defined as things a rational person wants whatever else he wants.[20]

Saying all this, however, leaves it as part of his project to find some basis for this index. Rawls thought that the problem was mitigated by his theory, since he had to deal only in broad categories like income and opportunities, and had to identify only whether the worst off were better off under one arrangement than under another. In fact, though, his theory ends up placing great demands on this index. What turns out to matter, in his theory, is the life prospects of those in the worst starting positions in life. To tote up uncertain prospects, we need more than an ordering from best outcomes to worst. We need to know what differences are big and what are small. We need what measurement theorists call a cardinal scale, and we need to be able to compare people on this scale. I am asking about the basis we might have for making these comparisons.[21]

As for Scanlon, he never, so far as I know, engages directly the kind of prospective argument that lies at the center of Rawls's "Original Position" and Harsanyi's theorems. I have followed Harsanyi and Rawls in saying how we can dismiss some objections as unreasonable. You do something and I object. You reply, "That's the established way we do things, and that's what you would have wanted before you knew which of us you would be and so how in particular it would affect you." This seems to show my objection unreasonable, and from crediting such dismissals Harsanyi draws powerful consequences. We are now seeing, though, that he must place great demands on the notion of individual benefit.

Scanlon offers an extended critique of the notion of welfare or a person's good, and this, I take it, constitutes his response to Harsanyi.[22] His critique doesn't rule out the possibility of dismissing objections as unrea-

[20] Rawls, *A Theory of Justice,* esp. 62, 92.

[21] Ibid., 64, 76–83, 93–98. I discuss problems of characterizing a representative person with the resources that Rawls allows himself in "Disparate Goods and Rawls' Difference Principle: A Social Choice Theoretic Treatment."

[22] Scanlon, "The Status of Well-Being"; Scanlon, *What We Owe,* chap. 3. One excuse I have for scrutinizing the ethical import of Harsanyi's two welfare theorems when Broome, in *Weighing Goods,* has given such a thorough treatment of the issues is that Scanlon's critique calls into question the notion of a person's good that is central to Broome's argument, and that Broome mostly takes as granted.

sonable on prospective grounds in the kind of way I've been discussing. It does, though, place on an advocate of such tests the burden of saying what notion of benefit can play this role. Scanlon allows that conceptions of how well off a person is might be tailored to play a role in a moral theory.[23] Clearly, though, from the way he himself develops his test, he doesn't think the test of what we would have wanted from behind a veil of ignorance plays anything like the broad and systematic role in a coherent moral theory that Harsanyi thought it played. Scanlon's critique of the concept of a person's good is a serious one, and I'll be particularly concerned to grapple with it.

I'll be arguing that we can't derive the needed notion of individual benefit directly from the preferences that people have, or even the preferences they would have in ideal conditions. Instead, forming a conception of benefit is part of ethical thinking, part of thinking how to live among other people. That fits a part of what Scanlon himself concludes,[24] but if the retort I've imagined always deflates a claim that an objection is reasonable, then thought of prospective benefit may have a much larger role in coherent ethical thinking than Scanlon gives it.

Preferences for Being Another

To illustrate and explore the problem, let's return to the simple toy case of Ida and Jay. I'll suppose first that we understand the notion of a person's good. Suppose again that how well off Ida and Jay would be in states *A* and *B* goes as follows:

STRUCTURE	*A*	*B*
Ida	9	5
Jay	1	3
EXPECTED	5	4

Harsanyi's argument favors structure *A.* Before they both knew who they would be, both would prefer structure *A* with expected level 5 to structure *B* with expected level 4. Jay comes out badly under structure *A,* but if *A* is the going structure, he has no reasonable objection to it.

[23] Scanlon, *What We Owe,* 110. Scanlon gives Rawls's primary social goods and Sen's capability sets as examples of such conceptions.

[24] Ibid., 110. It also fits Broome, *Weighing Goods,* 220.

I now turn to the objection raised by David Gauthier, Kotaro Suzumura, and others.[25] What do the numbers in my simple table represent? They represent evaluations from behind a veil of ignorance, made by people, as we imagine it, who are choosing a basic social structure only for their own prospective benefit. We are asking what notion of benefit feeds into the moral force of the rejoinder, "That's the structure you would have chosen for your own benefit." Ida, from behind the veil of ignorance, contemplates two possibilities: that she will be Ida and that she will be Jay. Now Ida in the world, let's suppose, wants a big funeral, and in state *A* she gets it, whereas in state *B* she doesn't. Does she have this preference behind the veil of ignorance? Suppose she does but that Jay doesn't. Jay understands well enough that, in case he turns out to be Ida, his actual strongest concerns include having a big funeral. But being Jay, he doesn't intrinsically care about having a big funeral. He is indifferent between being Ida and having a big funeral and being Ida and unexpectedly being cremated after death without ceremony. Being Ida, he understands, includes wanting a big funeral, but as he is, behind the veil, he is indifferent between (1) wanting a big funeral and getting it and (2) wanting a big funeral and not getting it. "If I get it," he figures, "I won't be around to enjoy it, and if I don't get it, I'll never miss it."

Once we distinguish Ida's preference for being Ida in *A* from Jay's preference for being Ida in *A,* we might get a more complex table like this:

Ida's preferences for being in state	*A*	*B*	Jay's preferences for being in state	*A*	*B*
as Ida	9	5	as Ida	5	5
as Jay	1	3	as Jay	1	3
Expected	5	4	*Expected*	3	4

Ida's evaluation of being Ida under structure *A* includes 4 units for having a big funeral and 5 for other aspects of how things go for her. Jay's evaluation of being Ida under structure *A* includes only the 5 units from those other aspects. From behind the veil, he places no value on actually having a big funeral in case he is Ida.

[25] Gauthier, "Refutation of Utilitarianism"; Broome, *Weighing Goods,* 55; Suzumura, "Interpersonal Comparisons." I discuss some of these issues in "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life."

Ida now can't refute Jay's objection by saying that *A* is the state he would have chosen if he hadn't known who he would be. The state he would have chosen is *B,* which gives him an expected utility of 4 as opposed to 3. Suppose, though, that structure *B* has been instituted, and Ida, not getting a big funeral, objects. (Or since she's not around to object, suppose someone objects on her behalf.) Jay can't deflate the objection by saying that *B* is the structure she would have chosen if she hadn't known who she would turn out to be. For each state there is an objection that can't be shown unreasonable, at least in this way. Unless we can find some other way to show one of the objections unreasonable, we're damned whichever state we institute.

That fits in with Scanlon's critique. There is no one coherent notion, he says, that will do the jobs that "welfare" or "a person's good" has been asked to do in much ethical thinking: roughly, determining (1) what a person will choose insofar as others aren't affected, (2) what others concerned to benefit him will choose to promote, and (3) what counts as the person's good for purposes of moral thinking.[26] We are asking about (3), and indeed, as Scanlon told us, not finding a way to read off a person's good from her preferences.

An appealing way out might be to let Ida be the judge of her own good. The problem remains, though—as Rawls insisted. From behind the veil of ignorance, in his system, we are choosing among alternative basic structures of society. What people will want, at base, might be highly affected by the kind of society we choose to have been nurtured in. Ida might have been indifferent to a big funeral if she had grown up in a different sort of society, in an alternative social order from among those that are open to choice from behind the veil of ignorance.

Rawls, as I say, was responding partly to this kind of problem when he set up his "index of primary social goods," but he offered, I think, no adequate, defensible rationale for its solution. I am asking whether such a rationale can be provided.

The Question of a Person's Good

Let's call the retort I've been discussing the "You'd have agreed" retort. This retort to an objection, recall, has two elements. First, "That's the way we do things." What you object to is a feature of our going practice. Second, "Before you knew how you in particular would turn out to be affected,

[26] Scanlon, *What We Owe,* 108–43.

you would have agreed to the practice—and for your own advantage." This retort does seem to have moral force.[27] Some notion of advantage and disadvantage, moreover, seems hard to escape in our moral thinking. Objections to a social order often are on the grounds that it disadvantages someone unfairly. Such an objection itself appeals to a notion of advantage or benefit, and so if the retort is incoherent because it requires a notion of a person's good, then so was the original objection. Still, we are left to ask what kind of force to accord a retort like this. The retort so far is vague; how can we spell it out in any precise way that will carry moral force?

Our question concerns the basic moral arrangements by which we live together. If we are to make sense of what we would have agreed to, we can't just look to our aims as they are as a result of the basic moral arrangements we have. The retort, if it is to have specific content, must be filled in with coherent fundamental aims we can take ourselves to have from a standpoint that doesn't just take us as we are. We must be able look to the various sorts of people we might have turned out to be under various different social circumstances, and ask how well these fundamental aims for oneself are fulfilled in these various kinds of lives. Will a plan for living with others, then, respect the "You'd have agreed" retort under some such interpretation?

In the rest of this lecture, I'll be considering one particular kind of way to work out the contractarian ideal, the ideal of living with others, if one can, on a basis that no one could reasonably reject. The way is to take the "You'd have agreed" retort and give it an interpretation suitable for answering fundamental moral questions. I won't settle what the interpretation should be. (I wish I could, but I can't.) Nor will I establish that this is the only fully coherent way to work out the idea of a basis for living together that no one could reasonably reject. What I'll be doing, rather, is to characterize a plan for living that incorporates such an interpretation of the ideal.

I plan to live with others, if I can, in mutual respect, on a basis that no one could reasonably reject on his own behalf. This plan constitutes an intuition on how to live with others, and as a plan, it can be couched as an imperative: "Prefer most to live with others on a basis that no one could reasonably reject on his own behalf." The intuition, though, is vague; crucial terms in the plan are left unexplained. We must specify what it is to

[27] Brian Barry, *Theories of Justice,* 334–35, questions the moral significance of Harsanyi's first welfare theorem, as does Broome quoting Barry in *Weighing Goods,* 56–57. To my own moral sensibility, the theorem has the moral significance that I am indicating.

reject a basis for living with each other *reasonably* and *on one's own behalf.* Harsanyi and Rawls offer an interpretation, a partial standard for what disqualifies an objection as unreasonable. In thinking how to live with each other, we may fill in our plan for living with each other with their proposal. Here is a partial interpretation of the indeterminate plan, a way to fill in the plan to live with others on a basis that no one could reasonably reject on his own behalf. "A rejection on one's own behalf of a going social arrangement is unreasonable if, absent information about which person one would turn out to be, one would have rationally chosen that arrangement on one's own behalf." This specification of one's plan for living with others, though, is still badly incomplete. It leaves to be explained choosing a thing "on one's own behalf" or for one's own sake. Uninformatively, we can put this in terms of a person's good. "One chooses rationally on one's own behalf only if one chooses what is prospectively most to one's good."

The Total Good of People

A plan that satisfies the three conditions I have stated will be a plan to maximize the total good of people. For suppose that one's plan satisfies these conditions, and consider a social arrangement that for each person, absent information about who he is, is most to his prospective good. Everyone would choose this arrangement on his own behalf, and so no one could reasonably object to it on his own behalf. A plan that satisfies these three conditions, then, will require living with others on this basis. Now for a fixed population, as Harsanyi's first welfare theorem showed, the basis for living with others that is most to one's prospective good behind a veil of ignorance is the basis that maximizes prospects for the total good of people. The plan that satisfies these three conditions, then, is a plan to maximize prospects for the sum total good of people.

The three conditions left us, though, with an uninterpreted term, the term "good" in the phrase "my good" or "your good." Scanlon's challenge is to find an interpretation of this notion of a person's good that lets it play a role in these axioms. What constitute preferences on one's own behalf? The requirement on such an interpretation is a planning requirement: we need an interpretation that goes some way to fill out how to live with each other on a basis of mutual respect. A person will be convinced of the interpretation if she plans to want most to live with others on a basis of mutual respect as so interpreted. I hope, then, to address people who, like me, plan vaguely to live with others on a basis of mutual respect if we can, and I follow Harsanyi and Rawls in proposing a form that such a plan might take.

The question for each of us is then whether to live in a way that takes this form.

This gives us a meaning for talk of a person's "good." A person's *good,* we can try saying, is whatever plays this role in the way to live. We accept that there is such a thing as a person's good when we restrict our plans for how to live with each other to ones that take the form displayed in the axioms. We accept some particular answer to the question "What is a person's good?" when we plan to live with others in a way that fits the axioms. What we regard as a person's "good" is then whatever plays the role of a person's "good" in the plan we have that fits those axioms. The interpretation we then accept is whatever interpretation of the axioms we plan to live by.

Notice, I have been speaking, so far, not of what really does constitute a person's good but of what it is to *accept* an answer to the question of what constitutes a person's good. The question of what constitutes a person's good is, I have been saying, a planning question. The meaning of a planning term can't be given a straight, naturalistic definition, in terms suited to empirical psychology. All we can say in straight terms is this: "A person's good is whatever it is, if anything, that figures in the way to live with others in a certain way. That way is specified by the three axioms. Whatever plays that role in the way to live, if anything does, is a person's good." What we can say further about the concept can only be oblique. We can say what it is for a person to *think* or *regard* something as constituting a person's good. To do so is to have a plan for living that takes the form of the axioms. What one then regards as a person's good is whatever plays the role given by the term "good" in those axioms.

The string of three conditions is a formal constraint on how to live with others. The constraint is to live with others on some specification or other of the ideal of fair reciprocity. Which specification of the ideal to adhere to is a further planning question, a further question of how to live with others.

If our preferences for how to live, as we struggle to make them coherent, do take this form, then we can go on to argue, using Harsanyi's first welfare theorem, that behind the veil of ignorance one would choose the social arrangement that, in prospect, maximizes the sum of people's goods. Preferring to live on that basis, one prefers to do one's part in an order that maximizes the total good of people, provided that everyone else can be fully expected to do so.

Is There Such a Thing as a Person's Good?

In this lecture I have been drawing on Harsanyi's first welfare theorem and applying it to interpret the appeal of Scanlon's talk of what no one could reasonably reject. The interpretation I proposed is one that Scanlon himself would repudiate, and nothing I have drawn from Harsanyi in this lecture shows him incoherent in this. It remains to be seen whether there is a coherent alternative to the kind of interpretation I have been proposing.

I hope, though, that I have given some glimmering of what speaks for Scanlon's reasonable rejection test when it is given an interpretation that takes this form. I interpreted the test as assessing rejections on one's own behalf. A rejection with moral force, my assumption was, must be a rejection on behalf of someone or other—and if it is on behalf of someone other than the person who does the rejecting, the question becomes whether that other person could reject the arrangement reasonably. The ideal, then, is to live with others, if one can, under an arrangement that everyone adheres to voluntarily, because it is an arrangement that no one could reasonably reject on his own behalf.

Talk of doing things on one's own behalf amounts to talk of doing them for one's own good as one sees it. Scanlon challenges traditional ways that ethical theorists have used the notion of a person's good, and so challenges the intelligibility of such talk. On the account I have given, the question of whether there is such a thing as a person's good is a planning question. It is a question of whether to live in a way that takes a certain form. I come to a view about what a person's good is, then, if and when I come to have preferences that take this form. We come to a joint view, in discussion, of what a person's good is if we all come to have preferences that take this form, and—crucially—for each of us the same valuations play the role these conditions assign to a person's good.

So far, this may well fit in with what Scanlon would expect. One of the functions that the notion of well-being has been meant to serve, he says, is "to be the basis on which an individual's interests are taken into account in moral argument." Moral principles will do such jobs, though, he thinks, with a variety of notions of a person's interests or good, and no single one of these notions will play the comprehensive moral role of being what the correct moral theory tells us in general to distribute.[28] I am now saying that *if* there is something that plays this comprehensive role, then that is what counts as a person's good. We are still left, though, with the question

[28] Scanlon, *What We Owe,* 136, 138–40.

of whether anything does. My own proposed interpretation of Scanlon's reasonable rejection test supposed that there is, but it is an interpretation he himself rejects, and I have not shown that it was the only possible coherent interpretation.

In the next lecture, I turn from Harsanyi's first welfare theorem to his second. I ask how it constrains the ideal social contract—the arrangement for living together, if any, that no one could reasonably reject. This theorem, I'll argue, can be interpreted in a way that makes it compelling, and in that form the theorem does sharply constrain what the ideal social contract could be.

III. COMMON GOALS AND THE IDEAL SOCIAL CONTRACT

In the first lecture, I proposed an account of what our job is in ethical theory. It is one of planning how to live with each other. Each of us plans how to live with others, and how to feel about things that he and others do or might do. With regard to planning, I cited a family of arguments from twentieth-century decision theory, the arguments of Ramsey, Savage, Hammond, and others. These arguments start with requirements of coherence in planning. They conclude that any ideally coherent planner in effect maximizes expected value on some scale. We could represent her plans, that is to say, by ascribing (1) numerical probabilities to eventualities and (2) numerical values to possible outcomes, and then evaluate each strategy for living by the values of the outcomes that it might have, each weighted by its probability. It has been controversial whether the conditions on plans that these arguments invoke are genuinely requirements of coherence, but I haven't seriously entered into those debates. Rather, I have been concerned with what follows if this tradition in decision theory is right.

In the second lecture, I cited two other major twentieth-century findings, Harsanyi's two welfare theorems. The theorems seem to show that the only coherent ethical theory is utilitarian in form. Utilitarians judge social arrangements by the total benefit they deliver. Specifically, Harsanyi's second welfare theorem placed three conditions on evaluating prospects: (1) evaluations of prospective individual benefit are coherent, (2) ethical evaluations of prospects are coherent, and (3) anything that is prospectively better for everyone is prospectively better ethically. Harsanyi showed, from these conditions, that if ethics treats everyone alike, then ethical value is a sum of individual benefits.

Equipped with this theorem, I took up the planning question of how to live with others—restricting myself to plans that place a premium on living with each other on a basis of mutual respect. I took up Scanlon's proposed interpretation of this standard: to live with each other on a basis that no one could reasonably reject. I explored how far the "You'd have agreed" retort could be taken, and this led to the aim that Harsanyi and Rawls propose. The aim is to live with others on a basis that we would have agreed to in ideally fair conditions, each with a view to his own prospective good—provided that this way of living together is the established way we do things. All this gives us at most a fragment of a plan for living with others, a plan for the case of "full compliance." It applies, that is to say, to the special case where our established ways of living together are the ones we would have chosen in fair conditions.

This interpretation of contractarianism, though, helps itself to talk of an individual's good. We must ask whether there is any conception of a person's good that makes the contractarian ideal, so interpreted, an ideal to plan for. If there is, then Harsanyi's first welfare theorem seems conclusive. If one's preferences in ideally fair conditions are coherent and one doesn't expect more to be one person than another, then one in effect values each outcome as the sum of the way one values it in case one is each of the people that, for all one knows, one is. At this point, however, enters Scanlon's critique: though loose talk of a person's good makes rough and ready sense, there's no one thing, he argues, that plays all the roles that have traditionally been ascribed to a person's good (or to welfare, utility, interest, benefit, or the like). I in effect accepted much of this critique. One role that Scanlon does allow to the notion of a person's good or interests, however, is that of counting in a particular way for particular moral purposes. (An example is Rawls's index of "primary social goods" such as money, powers, and opportunities.) As Scanlon himself works out of his "contractualism," no highly general notion of a person's good or interests plays any comprehensive role. I am asking whether Scanlon is right about this. In particular, do Harsanyi's welfare theorems compel us to develop a conception of a person's good or interests and then conclude that morality consists in promoting a general interest—a value composed of individual interests? Is it incoherent to think otherwise, once we think that morality is made for humanity?

The main point of the second lecture was still to ask *about* the questions we are asking. I looked at two questions: First, is there any such coherent thing as a person's good? Second, if so, what is it? What is a person's good? These both, I said, are planning questions. We interpret talk of "person *i*'s good" when we say what form a person's preferences must take for him to think that there is such a thing, and have an opinion as to what a person's good is. I thus characterized, in indirect terms, what a person's good is if there is any such thing.

If we start out taking the concept of a person's good or benefit as intelligible, then Harsanyi's second welfare theorem, even more than the first, makes it hard to see how Scanlon's reasonable-rejection test could lead to anything but agreeing to maximize the total prospective good of persons. We would reasonably reject a social arrangement if it is wasteful, if some alternative would give us each a greater prospective benefit. Our conception of individual social benefit is presumably coherent. As for prospective ethical value, I'll discuss that briefly later, but suppose for now that we would agree to a coherent conception of value from the ethical point of view. That gives us all the conditions of Harsanyi's second welfare theorem. If we treat everyone's good alike, the theorem then says, we agree to maximize the total good of everyone.

What now, though, if the very notion of a person's good is in question? Still, I'll argue in this lecture, Harsanyi's second welfare theorem (or something close to it) tells us the form that a coherent social contract will take—its formal structure. Doing this leaves open the question of how to fill in the structure. Harsanyi's second welfare theorem, like the first, is in part an abstract mathematical result, which can be given various interpretations. Harsanyi had his own interpretation, but even if the assumptions of the theorem don't all hold under that interpretation, they might all hold under another. Both theorems are mathematically correct, and so the debate must be over whether any interpretation of these mathematical results is of ethical import. Specifically, is there any interpretation under which Harsanyi's second welfare theorem shows that a coherent ethics must take something like a utilitarian form?

Much of ethical theory, over the past few decades, has been devoted to showing that there are things to care about and to want others to care about, in living with each other, that don't take the form of summing up the good of individuals, under any conception of what a person's good consists in. Each person has special concerns and responsibilities, and shouldn't be expected just to place them on a par with the concerns and responsibilities of everyone else. The Jeske and Fumerton canoe example was meant to give vivid intuitive support to such a picture of the demands of morality. Harsanyi's second welfare theorem, though, I'll be arguing, shows that this antiutilitarian picture won't fit in with contractarian thinking.

I am taking it, remember, that Peter Hammond's argument, or another like it, establishes that requirements of decision-theoretic coherence apply to the totality of aims that a person has reason to advance, the totality of considerations for a person to weigh in making his decisions. It doesn't immediately follow that there is such a thing as the self-interested component of those aims, and Scanlon may be denying that it follows at all. I will argue that it does follow—but my argument will be indirect.

The Kingdom of Ends

By the *ideal social contract,* I mean the way of living together that no one could reasonably reject. (I'll ignore the question of whether there is such a way or whether there might be more than one such way.) Suppose, then, for the sake of inquiry, that Scanlon is right, and the ideal social contract doesn't take the form of settling what is to count as a person's good, and then agreeing each to advance the sum of everyone's good. What possibilities does that leave open?

Here is a first question about the ideal social contract: The contract places constraints on the ways each of us is to pursue his aims. These constraints must be ones that it is rational for each of us to abide by, given that this particular social contract spells out our established ways of living with each other, and given the rationality of wanting to live together on a basis of mutual respect—interpreted as living in ways that no one could reasonably reject. Suppose, then, each of us acts rationally and abides by those constraints. Since we abide by the constraints rationally and voluntarily, our plan of action, in light of this contract's being in force, is coherent. That entails, we are supposing, that it satisfies the Hammond conditions, and amounts to maximizing expected value on some scale. Here, then, is the question: are we all, under the ideal social contract, to have a common set of aims? Does the agreement we would have arrived at, in ideally fair conditions, take the form of agreeing to a common set of aims—aims that somehow accommodate what each of us has reason to want in life? Would our agreement be each to maximize expected value on the same scale? (If so, then what's up for negotiation in arriving at the social contract is what this common scale is to be.) Or alternatively, would our agreement allow each of us to pursue her own set of aims, different from the aims of others but somehow constrained to accommodate things that others have reason to want?

We are asking about what Kant dubbed the "kingdom of ends." On the predominant interpretation of Kant, the kingdom of ends is an arrangement that each of us wills, whereby we can each pursue our separate ends in a way that duly accommodates the ends of others. This reading fits much of what Kant says. An alternative, though, would be to conceive the kingdom of ends in a more utilitarian way, with each of us accommodating the ends of others by incorporating them into her own aims, weighing the ends of each person equally in her decisions. She still pursues her own ends, in that her ends count in equally with everyone else's. Others too count her ends equally with theirs—but normally, of course, she is in the best position to advance her own ends. Clearly, Kant rejected this as what he meant by the kingdom of ends, but the question remains whether any other systematic sense can be given to the ideal.[1]

Now as an interpretation of the ideal social contract, the first alternative, I'll argue—allowing us each to pursue a different set of aims—is incoherent. Suppose the ideal social contract did take this form. Each of us, we have agreed, is free to have various aims that satisfy the conditions of our agreement, different from the aims that are to guide the decisions of others. We each adopt such a separate set of goals, suppose. Since we act rationally in doing so, the goals can be represented as a scale of value to be pursued. Call this the person's *goal-scale.* My goal-scale, note, doesn't then represent just my own good in any normal sense of the term. It makes some accommodation of my ends to the ends of others—to their good, or to other things they have reason to want. The scale presumably puts great weight, for instance, on not killing you, even if I could get away with it and even if killing you would greatly advance things I have reason to want. My goal-scale thus accommodates your end of not being murdered, whether that end is to my own good or not.[2] My interests, in some sense, will figure into my goal-scale, but they won't be all that determines it—and my interests figure somehow into the goal-scales of others too. That is the sort of thing that, on this conception, an ideal social contract would require.

Now the problem for such a social contract is that diverging goal-scales can make for prisoner's dilemmas. That is to say, there will be cases where

[1] Kant, *Grundlegung* (1785). R. M. Hare in "Could Kant Have Been a Utilitarian?" argues that although Kant was convinced that his system yielded the pietistic morality of ordinary people of goodwill, his system cannot be made to yield the results he wanted except by making unsupportable and ad hoc moves. Most other recent and current Kantians think that a Kantian rationale can be given for a morality that departs fundamentally from utilitarianism.

[2] We could instead use the term "utility scale" for what I am calling a goal-scale, and thus latch on to one of the meanings that highly theoretical economists and decision theorists have for the term "utility": a scale representing, in a canonical way, how a person is disposed to make his decisions. The term "my utility," though, also suggests my good or my interest, and we must sharply distinguish the scale I adopt under the terms of the social contract to guide my choices from my own good or my own interests, which the social contract accommodates.

one prospect, *X,* comes out higher on everyone's goal-scale than does another prospect, *Y,* but where if each of us guides his choices by his own goal-scale, we will end up with prospect *Y.* We could, in such a case, have agreed on a shared goal-scale that would end us up with *X*. Thus, whatever is to be said from my point of view for coming higher on my goal-scale, and whatever is to be said from your point of view for coming higher on your goal-scale, there's more to be said from both our points of view for *X* than for *Y*—yet the social contract tells us to act in ways that combine to achieve *Y.* This seems an incoherent way to arrange our lives, a way with no intelligible rationale. Any of us can reasonably reject the arrangement as wasteful of that which is worth his pursuit.

The work here is being done by Harsanyi's second welfare theorem under a new interpretation—or more precisely, by a variant of the theorem. Consider first the original theorem on this new reading: an individual's prospects we now read as his goal-scale, the scale on which he acts, in light of the social contract, to maximize prospects. Harsanyi's first condition thus becomes simply that each individual has a coherent policy for action, representable by a goal-scale. The second condition of the theorem, the prospective Pareto condition, we now read as ruling out a social arrangement if some alternative comes higher on everyone's goal-scale. The third condition is now that social policy be coherent.

This third condition, though, is open to question, and handling this issue requires not precisely Harsanyi's theorem but a variant. For our purposes, it turns out, we can drop the third condition. Consider all the prospects we could jointly bring into being by each adopting a complete contingency plan for action. Consider any one of those prospects that satisfies the prospective Pareto condition. There will be a goal-scale that is a weighted average of individuals' goal-scales for which this prospect comes highest.[3] Thus, we can argue, if individuals abiding by the social contract have distinct goal-scales, and if collectively their policies for action yield a prospect that satisfies the prospective Pareto condition in terms of their respective goal-scales, then there is a possible goal-scale such that if they adopted it as their common goal-scale, they would reach this outcome.

I'm not now appealing to any suspect notion of a person's good. Even if there is such a thing as a person's good, his goal-scale, as I have said, represents not just his own good. Rather, it reflects all that he has reason to aim for, given that the established ways of doing things accord with an ideal contract, and given that he has reason to abide by this established social

[3] For a more precise formulation, see the Appendix.

contract voluntarily. I am not now assuming that, in agreeing on the social contract, each of us would be concerned solely to advance his own good. I'm appealing, rather, to an incoherence in the rationale for any social contract that allows us to pursue goals that might conflict.

I began, in the first two lectures, with schematic cases of children needing rescue. These weren't cases, note, where it is clear what constitutes a father's good. The grounds we recognize for a father to have special concern for his own children aren't just a matter of the gratification he gets from them and the anguish of losing them, but of special parental responsibilities. Indeed, if we ask what component of a parent's concern for a child is self-interested, the question may have no clear sense. Still, as we saw, whatever special reasons a father has to want his own children in particular not to drown—reasons he doesn't share with fathers of other children—those aims may be better advanced in prospect by a social contract that tells us each to weigh the safety of others' children as he does the safety of his own. I am now saying that this lesson generalizes. Any social arrangement that lets us pursue divergent goals suffers a like incoherence. Whatever reasons each has for the peculiarities of her own goals, there is a way better to advance, in prospect, all those goals at once. The way is to agree on a common scale of goals for all to pursue. The way is to agree, as we might put it, on what to treat as the overall good, and then for each of us to advance the overall good as much as possible.

By *the overall good,* then, I shall mean good as measured by whatever goal-scale would be specified by the ideal social contract. It is whatever goal-scale it would be unreasonable for anyone to reject as the goal-scale for us each to take as his own. The scale that gauges the overall good is the one we would agree to use, in effect, to guide our decisions. We would agree always to do whatever offers the best prospects as measured by that scale. We would agree, that is to say, to do whatever maximizes the rationally expected value of the overall good.

The Common Ends to Adopt

A social contract with a coherent rationale, I have been arguing, will designate a goal-scale for us to adopt in common. What I'm to advance, you too are to advance. But what will this common goal-scale consist in? It must somehow take all of us into account. Morality, after all, is made for humanity, not the other way around. If a person is reasonably to reject a proposed arrangement, it must be on the basis of something a person has reason to want from a social contract. If this isn't the person's own good, or if there isn't any such definite thing as a person's own good, the basis must still be something worth wanting—worth wanting from that person's own standpoint and on grounds that don't invoke preconceived demands of morality.

To say all this, though, is not to specify just how the overall good takes us into account. What is this overall good to consist in? This question, if what I have been saying is right, is a planning question, a question of what to want from a social contract. A crucial part of ethical theory will be to discern a basis for adopting some particular common goal-scale. This planning question is one that I haven't yet addressed. In particular, I haven't derived, from Harsanyi's second welfare theorem, that the overall good adds up everyone's individual good. I am not even assuming, at this point in the argument, that there's any sense to be made of talk of a person's individual good. Indeed, from the austere materials I am allowing myself, I won't be able to derive such a conclusion. Decision-theoretic requirements of coherence in action won't by themselves entail that the common goal for each of us to pursue, in living together on a basis of mutual respect, adds up, in any sense, the good of each of us. Perhaps the overall good is formed in this way, but I won't be able to demonstrate that it is.

Here, though, is something that does follow from requirements of coherence. Take any consideration that weighs into the overall good. For all we have said, some of these considerations may concern no one in particular. Perhaps, as part of the social contract, we are to promote diversity of species on the planet. It is to count in favor of an action on the part of anyone, we might agree, that the action would promote species diversity. (I'm not discussing here whether species diversity indeed is something to promote for its own sake, just saying that coherence requirements don't rule this out.) Such a common goal, we can say, is *impersonal,* as opposed to *person based.* With other goals that we are to take up in common, the grounds for doing so involve, in one way or another, an individual. They are considerations for the rest of us to weigh, under the social contract, because of the way their relation to that person gives her reason to want us to weigh them. Suffering presumably has this status: your suffering pertains to you, and it is because you have reason to want not to suffer and so to want the social contract to work against your suffering that the social contract will tell everyone to want you not to suffer. (More precisely, it will tell everyone to treat the fact that you would suffer if something were done as weighing against doing it.) Now suffering is bad for a person if anything is, but other things that people tend to want have a status that is less clear. Prestige, honor, recognition after death, integrity, family thriving—these things are puzzling. If, though, the social contract tells us to give intrinsic weight to any of these, the grounds will presumably be person based.

Suppose, then, a consideration has ethical import; the social contract tells us each to weigh it. We can ask whether the import is person based or impersonal. Coherence doesn't demand that it must be person based, for anything we have established, but if it is person based, that gives the consideration a status worth singling out. A *person-based* consideration we can define as a consideration pertaining to some specific person that has moral weight because of how it pertains to him, and because of how the way it pertains to him gives him, in particular, reason to want it fostered by the social contract.

It is probably best, at this point, not to speak of a person's "good" but of his *interests.* (Scanlon adopts this usage.)[4] Our question now, after all, is not directly what to want in life for one's own sake but what to include in the social contract, what considerations to agree to give weight to. The argument I have given doesn't establish that the person-based considerations to promote under the social contract must count as aspects of the person's "good" as we normally use the term. One interpretation we might now give to talk of a person's "interests" is this: a person's interests consist of those things that are of ethical import because, in this sense, they are *based* in him.

Trivially, any consideration that bears on the overall good is person based or not; if not, it counts as impersonal. Suppose, then, there are no impersonal goods, that every consideration that the ideal social contract tells us to take into account is person based. Will it follow that the overall good is the sum of individuals' interests? To establish this, we need one further assumption: that the common goal-scale that the contract prescribes—the scale that measures the overall good—sums up the weights of a set of considerations. Given this, since each consideration must either be person based or impersonal and none of them are impersonal, they must all be person based. The overall good, then, is measured on a scale that adds up the weights of person-based considerations—which is to say, of individuals' interests. Now I find it hard to see how a coherent goal-scale can have any rationale other than that it sums up the weight of a set of considerations. I don't know how to establish definitively that it must, but in the rest of what I say in this lecture, I'll assume that it must. If it does, the argument I have given shows that the overall good is composed of the interests of individuals.

[4] Scanlon uses the term "interests" in this way; see, for instance, *What We Owe,* 136.

All this assumed that there are no impersonal goods. Suppose instead that there are such goods. (Pick your favorite candidate; my example was species diversity.) Then by the same argument (with the same additional assumption), the overall good is composed of individual interests plus whatever impersonal goods the ideal social contract would include.

In either case, then, the social contract will tell us each to adopt a common goal-scale, and this goal-scale will be the resultant of our individual interests—along with, conceivably, certain impersonal goods.

What Is in a Person's Interests?

Consider three questions: (1) Is there such a thing as a person's interests? (2) If so, what are they? (3) Will the ideal social contract tell us each to pursue the sum of individuals' interests? I have been asking what these questions mean, and the meaning that we can give to talk of a person's "interests" on the basis of what I have been saying is this: a person's interests consist in whatever has moral weight because of how it pertains to her and how the way it pertains to her gives her in particular reason to want it fostered by the social contract. The three questions, as I'll interpret them, are all planning questions, questions of how to live with others. Harsanyi's second welfare theorem determines answers to transformed questions (1) and (3). It determines answers, that is, supposing that there is a basis for living with each other that no one could reasonably reject. This way of living together—the ideal social contract—is for each of us to adopt the same goal-scale, a scale that somehow accommodates things each of us has reason to want. Whatever this scale measures I call the overall good, and the way it is composed settles what counts as a person's interests. Thus, (1) There is such a thing as a person's interests, and (3) The ideal social contract says to pursue the overall good, composed of the interests of all people plus, conceivably, of impersonal good. (This assumes, remember, that the overall good is a resultant of considerations.)

That leaves question (2). What is in a person's interests? I have said that this is a planning question. It is roughly the question of what to count in the social contract. I haven't, though, addressed this question. It is one of the questions in ethics that I would like most to answer, but not a question that I aspired to answer in these lectures. I have been interpreting the question and asking what form a coherent answer must take. Trying to answer the question must lead to almost all the questions that ethical inquiry addresses.

Let me speak briefly about this question, though. Hedonic goods obviously enter in: happiness, enjoying what one does and getting satisfaction from it, freedom from suffering, positive "hedonic tone," and the like. One chief puzzle, debated over the years, concerns what used to be called "ideal" goods. G. E. Moore listed as the greatest goods "organic wholes" involving pleasures of friendship and pleasures of contemplating beauty.[5] These things involve pleasure and more: roughly, that one derives pleasure, in characteristic ways, from genuine friendship and genuine beauty. James Griffin ventures a list of prudential values beyond enjoyment as accomplishment, understanding, deep personal relations, and components of human existence such as autonomy, basic capabilities, and liberty.[6] One question, widely debated, is whether these really are things to want for their own sakes. Or are they instead things to want just because they reliably go with the greatest of pleasures? Either way, they are things to want in life and things to want our social arrangements to foster. Do they count intrinsically, though, as parts of what the overall good consists in? If they are worth wanting for their own sake, and if the ideal social contract tells us each to advance some common overall good, aren't these things worth counting among the things we agree to foster jointly?

Rawls himself thought not. He thought that not even enjoyment and freedom from suffering would figure among the "primary social goods" used to assess possible social arrangements. Some arguments against maximizing the total pleasure or happiness of people strike me as bad. (Nozick worried about "utility monsters" who make themselves highly sensitive to income level and the like, so that their needs for income will count more heavily under the social contract, and they will be awarded the lion's share of resources.[7] But a wise implementation of the social contract will heed incentive effects, and not reward a person's setting himself up to suffer unless rich. He may threaten to hold his breath until given a million dollars, but a wise system won't respond even if the threat is credible.) Other arguments for Rawls's position, though, call for careful thought; they concern what people's interests are and how they can be compared. An interrelated

[5] Moore, *Principia Ethica.* Moore didn't think these to be person-based goods in my sense; he treated all goods as impersonal. I take this to be far from the spirit of contractarianism, the kind of moral vision that I pursue in these lectures.

[6] Griffin, *Well-Being: Its Meaning, Measurement, and Moral Importance,* 67. Griffin's list is of course meant as rough and tentative.

[7] Nozick, *Anarchy, State, and Utopia,* 41. My reply here concerns a threat to suffer terribly if one isn't given resourses. It is true that if a person can achieve extraordinary happiness with additional money and not otherwise, utilitarianism will treat it as urgent for him to have the money. But wealth beyond dreams, we find, doesn't make for happiness beyond dreams; we can't make ourselves into "utility monsters" of this kind.

problem is how people are to live together on a basis of mutual respect when they disagree fundamentally about what to want in life and on facts that bear on how to pursue it. I'll try to lay out these problems within the metaethical framework that I have sketched in these lectures.

Interpersonal Comparisons and Reasonable Disagreement

Here is a first problem, serious but solvable: Begin with the question of what to want from a social contract on *self-based* grounds. By these I'll mean grounds that, because of how they pertain to oneself in particular, give one reason to want the social contract to accord them weight. What fundamentally self-based grounds there are is a planning question, and intelligible enough. What, then, constitute a person's interests? An interest of his, recall, we define as a consideration pertaining to him that the ideal social contract accords weight because of how it pertains to him, and because of how the way it pertains to him gives him, in particular, reason to want it fostered by the social contract. If the social contract is made for him and the rest of humanity, then it may seem that his interests in this sense are just the things for him to want from the social contract on self-based grounds.

How, though, if that is so, are we to compare the strengths of interests of different people? How is the social contract to trade off their interests against each other, when those interests can't all jointly be catered to? A person's interests may well depend on his personal characteristics, since what to want from a social contract might differ from person to person, in a way that depends on those characteristics. In saying this we must keep in mind the difference between two related questions: the psychological question of what the person *does* want and the planning question of what *to* want in case one is that person with that person's characteristics. Characteristics in which we differ may well matter for both, but here, remember, our question is what *to* want. We differ in what gives us a sense of meaning and fulfillment in our lives, we differ in our ideals for ourselves, and so there may be different things to want in case one is like you and in case one is like me—or things to want in different strengths. You thrive on controversy, perhaps, and I on dogma, and we can protect my sensibilities or give scope to your free tongue and spirit. Protection is something to want in case one is like me, imagine, and scope in case one is like you. How are we to compare, then, the urgency of protecting me and of giving you scope, when we can't do both. That is the first problem.

So posed, the problem seems solvable, in somewhat the way that

Harsanyi envisaged.[8] I can distinguish what to want from the social contract in case one is like you and what to want from it in case one is like me. I can compare how strongly to want things by facing the hypothetical planning questions of what to prefer if one is equally likely to be like you or like me, and choosing between being provided for in the one case and in the other. This gives us a comparison of person-based interests, and a person could reasonably reject, it seems to me, having his interests weighed at less than what this comparison indicates.

Rawls's aim, though, was to find a basis for living on a basis of mutual respect that suits people who disagree fundamentally on what to want in life. My discussion so far has assumed that planning questions like these have right answers, and that the terms of the ideal social contract can depend on what these right answers are. Whether planning questions like these do have right answers, answers that are interpersonally valid, is a difficult issue that I won't here try to address. (I struggle with this in both my books.)[9] Even if these questions do have right answers, though, we surely don't agree on them. Rawls's problem, couched in my terms, was how to live together on a basis of mutual respect in the face of perennial, fundamental disagreement on basic questions of planning and of fact. We are back to Ida who wants a big funeral and Jay who thinks that a big funeral is not a thing to want for its own sake even if one is like Ida. How can they live together in mutual respect, on a basis that neither would reject even if she had the power to force an alternative on the other?

Rawls proposed marking off a range of answers to basic questions of planning and of fact as "reasonable." His question was not the general one of how to live together with others on a basis of mutual respect in the face of any fundamental, perennial disagreement whatsoever. Rather, it was how to do so when others' views, even if mistaken, are reasonable. As for what counts as "reasonable," that must amount to a planning question. To treat a view as reasonable in this sense, we might try saying, is to be willing to accommodate it. It is to want to live with those who hold the view on a basis that they can accept, with a rationale that prescinds from questions on which they don't share the truth as one sees it. It is to prefer this to the alternative of imposing a social order on them—even for the case of having the power to suppress them.

This lecture has centered on a Harsanyi-like argument that what's up

[8] Harsanyi, "Cardinal Welfare."

[9] Gibbard, *Wise Choices,* 153–203; Gibbard, *Thinking How to Live,* 268–87.

for negotiation in arranging a social contract is what to count as a person's interests—and possibly what to count as impersonal goods. To be coherent, I argued, a social contract must specify a single goal-scale for us each to make his own. How, then, if the argument I gave is good, does it bear on Rawls's project? Even if the argument is good, it may still be that some people won't accept it, even when offered careful explanations. It may nevertheless be best to undertake to live with them on some basis that, given their views, they will accept. If "reasonable" views are ones to be willing to accommodate in a scheme of voluntary social cooperation, this means that even if everyone ought to accept the arguments I have given, rejecting them may count as reasonable. Moreover, suppose that everyone does accept the Harsanyi-like argument that I gave. Still, even if we all accept the same conception of a person's interests and all accept that we each are to advance the combined interests of everyone, we may disagree fundamentally on the facts that bear on how to do this. What we may still all be able to agree on, in that case, is a basic structure of society, a way to proceed in face of our disagreements—even though none of us thinks that it is the structure that most fosters the totality of people's interests.

The arguments I have given, then, speak to Rawls's problem only in special circumstances, circumstances where there is more agreement on what matters in life than Rawls envisaged. My arguments don't tell us how Ida and Jay are to live in a scheme of voluntary social cooperation and mutual respect when they can't agree on the worth that a big funeral would have for Ida once she is dead. Perhaps the two can agree to count a big funeral as in one's interests if one cared when alive and not if one didn't. Perhaps if they *can* so agree, then whoever is right on the worth of funerals, they each *ought* to so agree. Probably, they ought to agree on a scheme much like the one that Rawls advocates, establishing a basic social structure that gives both of them good prospects for multipurpose means like income and opportunities. Ida can then have her funeral if she or others want to bear the costs. No ethical conclusions along lines like these follow from anything I have said, and nothing I have said tells us how to choose among alternative economic schemes that share these overall features. We are left with the problem of how to compare different people's interests.

I won't finish these lectures with a solution to Rawls's problem—I'd love to, but I can't. Rawls himself, in my judgment, didn't come up with a compelling solution, and neither has anyone else. What terms of social cooperation are worth accepting when we disagree fundamentally on basic questions of fact and value must depend, I would think, on many questions of fact, psychological and sociological, and on the difficult question of what to prefer in light of those facts. What is the range of views to be reconciled, in what numbers? What are the effects of forcing people on various matters against their convictions? How are we to deal with these facts, and what attitudes are we to have toward the people with whom we disagree? It would be surprising if some game-theoretic scheme held a straightforward answer to such questions—though game-theoretic insights may be highly relevant.

Through all this indeterminate discussion of Rawls's project, though, the force of the Harsanyi-like result of this lecture remains. Any social contract will be self-frustrating unless it takes a certain utilitarian-like form: agreeing to maximize prospects on some common goal-scale. Otherwise, there will be a possible alternative social arrangement that, for each person, offers better prospects in terms of what that person is trying to accomplish. The lesson of the canoe examples survives. That leaves the question of how the common goal-scale of the ideal social contract is to be set. What if we disagreed fundamentally on the importance of saving our children or on how to assess the facts that bear on how best to save them? Confronted with the problem that Rawls sets, I have found no compelling, tractable answer. The finding remains, though, that without such a common-goal scale, our social arrangements are jointly self-frustrating.

Harsanyi and Beyond

If a social arrangement is jointly self-frustrating, I have been supposing, then anyone could reasonably reject it. Some alternative to it, after all, is better with respect to each person—better, that is to say, as reckoned in terms of the values that this very arrangement tells her to promote. Getting further in our inquiry, though, would require more examination of this claim: if a social arrangement is jointly self-frustrating, does that truly make it reasonable to reject the arrangement? Progress would also require careful scrutiny of other assumptions I invoked. Do those assumptions apply to our actual circumstances? Or do they at least apply to circumstances that are relevant to moral argument? What is the upshot when they don't?

Recall how my argument went. The conclusion was that a social contract must establish a common goal-scale for all of us to advance—on pain of being jointly self-frustrating. Start first with an individual: for him, sheer coherence in action requires pursuing some goal-scale or other; he will act as if he were maximizing prospects as reckoned by that scale. This is the upshot of arguments by Hammond and others; in these lectures I accepted those arguments with little scrutiny. Turn now to society and the social contract. If each individual pursues a distinct goal-scale—favoring, say, his own children or his own integrity—the result, it turns out, must be collectively self-frustrating. It will be self-frustrating in this sense: there will be an alternative goal-scale that we all might have pursued in common, thereby achieving prospects that are better on each person's goal-scale. This is the Harsanyi-like result that I have been exploring.

What rationale, I next asked, could there be for agreeing to a social contract with this blemish? Couldn't any of us reasonably reject such a self-frustrating social contract? For each of us, after all, what it tells him to pursue is better attained, in prospect, by the same alternative social contract, an alternative that hands us a common array of goals to pursue. This result goes partway to what utilitarians have always claimed, and it is at odds with many of our intuitive judgments—as with the canoe-rescue case of Jeske and Fumerton. If one's children are so greatly worth saving, the point is, why agree to less than best prospects for their being saved?

I can't claim, though, that such a challenge is unanswerable. Indeed, many cogent answers have been explored by ethical theorists. The argument depends, of course, on supposing that there is a form of social cooperation that no one could reasonably reject. This amounts to supposing that ethics, in a contractarian vein, is possible. I have chiefly assumed as well, implicitly, that we have an agreed basis for judgments of nonethical fact, a basis that lets us speak of "prospects" as reckoned by some particular goal-scale. I have assumed, moreover, that we would each implement whatever we agreed to, and implement it costlessly and with perfect rationality—and that we all know that we would. It was on these assumptions at least that I based my conclusions, and further inquiry demands seeing how those assumptions should be relaxed and what the upshot then is.

Compliance and its costs raise acute problems, in life and in moral theory. A social ethos never gets perfect compliance, and only costly efforts could achieve even partial compliance. Utilitarians have long faced this problem; they have proposed solutions for it and debated the adequacy of those solutions.[10] Rule utilitarianism and other forms of indirect utilitari-

[10] Sidgwick confronts these problems and sticks with a direct utilitarianism. Richard Brandt develops a form of rule utilitarianism in "Toward a Credible Form of Utilitarianism," "Some Merits of One Form of Rule Utilitarianism," and *A Theory of the Good and the Right.* Harsanyi, in "Morality and the Theory of Rational Behavior," endorsed Brandt's rule utilitarianism. R. M. Hare, in *Moral Thinking: Its Levels, Method, and Point,* distinguishes questions on two levels, the "critical" and the "intuitive." He argues that at the critical level, only act-utilitarianism fits the logic of moral thinking, whereas at the intuitive level of thinking we should accept precepts that are not directly utilitarian.

anism distinguish a background rationale for morality, which is utilitarian, with morality made for humanity, from the moral code for a society that the rationale supports. The moral code may invoke a notion of individuals' interests or their good and still not tell each of us to promote the total good or interest of all of us—even if the background rationale for the code is to promote the totality of people's "interests" in another sense of the term. There may well be, then, an indirect rationale for the sort of limited notion of interests that Rawls and Scanlon advocate, and for a moral code that tells us to heed not only people's interests but also their rights and their autonomy. Rawls too distinguishes a background rationale from the principles of justice that are to govern the basic structure of society, and he too expects only partial compliance.[11] A contractarian like Rawls must worry about partial compliance for a further reason that I have mentioned: his basic rationale for heeding morality is reciprocity, and with partial compliance, though there's something to reciprocate, there's less than there might be. In brief, no social contract we could want will draw full compliance, and the upshot is a central problem for ethical theory.

My discussion has left out much else that needs study, including much that is already receiving valuable treatment in the literature of moral philosophy. Among other things, I have said nothing about a person's right to make his own mistakes, about possible conflicts between respect for his autonomy and concern for his welfare. I think that much could be said about this and other matters within the kind of framework that Harsanyi sets up, but I have not myself been doing the work in these lectures.

The lesson I draw from Harsanyi, then, is crucial but limited. Any ethic that lets us pursue basically different purposes faces a challenge. The challenge does not end discussion, but it should inform any broad inquiry into ethical theory. Is it ethically permissible for any of us to give special heed to our own special concerns and our own particular responsibilities? Doubtless yes, but why? Why shouldn't any such claim be rejected as self-frustrating? We haven't established that an ethical theorist who makes

[11] By Rawls's background rationale, I mean his specification of the "original position" and his arguments that the test of principles of justice is what would be chosen in the original position as he specifies it. Roughly, in *A Theory of Justice,* the rationale for principles of justice is that we would have chosen them in a fair situation to govern the basic structure of society. In subsequent work, he expands on the "Kantian interpretation" of his theory, and stresses that the rationale involves expressing our nature as free and rational beings. Parties in the Original Position expect partial compliance in that although they know that the principles they choose will govern the basic structure of their society and be widely accepted, they do not expect unanimous acceptance of the principles or invariable conformity to them. They must provide for education and enforcement, and choose principles that, once implemented, would continue to be widely accepted and adhered to.

such a claim has no answer, but he does owe us one. What is the point of moral strictures? If the background rationale for the strictures isn't some aim we could have in common to accommodate our various individual ends, can the rationale be fully coherent?

Questions of ethics are, in effect, planning questions, I started out saying. They are questions of how to live with other people who face like questions. I have been addressing a range of fundamental questions of ethical theory as planning questions. The way to live with other people if one can, I took it, is on a voluntary basis that no one could reasonably reject. In accepting such an ideal for living together, I had to rely on intuitions on how to want to live with others. A crucial range of ethical puzzles then became questions of what to want from a social contract, and what sort of social contract to respect if it is in force. Requirements of coherence on plans, I began to argue, generate restrictions on the kind of social contract that no one could reasonably reject.

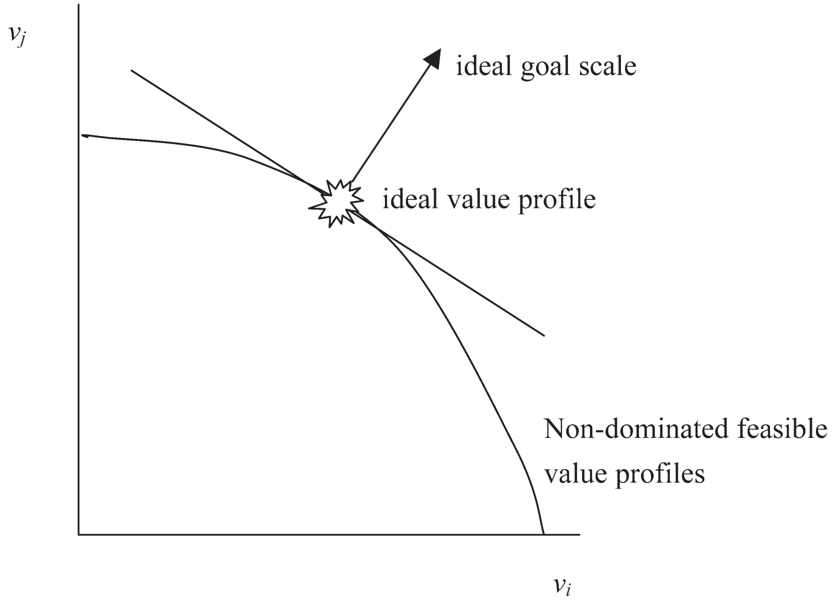
The demands of coherence in our ethical thinking can be powerful, and they sometimes run counter to strong intuitions. Amartya Sen and Bernard Williams published, almost two decades ago, a collection of articles that included both a lucid summary by Harsanyi of his ethical thinking and Scanlon's own initial exposition of his "contractualism." The editors entitled their collection *Utilitarianism and Beyond.* The title was apt in a way: we do still need to get beyond the point that ethical theory has reached. To do so, though, we can't move beyond utilitarianism and drop it. We must still heed the force of the kinds of considerations that Harsanyi raised. Any moral vision that doesn't specify a common goal-scale as a basis of its rationale must explain why it doesn't fall in the face of a Harsanyi-like result. We can forge beyond Harsanyi only by keeping careful track of what he showed.

Appendix. The Harsanyi-like Result

The "Harsanyi-like result" that I rely on in the third lecture is just the following. It is a form of argument familiar to economic theorists, although the niceties of just when it holds require some care. Start with all the possible policies for action that each person could adopt. A policy—or *strategy,* as I'll say to fit game-theoretic terminology—assigns an action to each informational state that one might be in. For each person, some conceivable strategies are feasible for him and others are not. Call an assignment of a strategy to each person a *strategy profile.* A strategy profile is feasible just in case each person's strategy for that profile is feasible for him. Assume a unique prior subjective probability measure that everyone shares at the start and then updates with new information. Then we can speak of the *prospect* that a strategy profile presents; it assigns to each possible outcome the probability that outcome would have if each person acted on his strategy for that profile.

Let each person have a goal-scale. Any prospect has an expected value on a given person's goal-scale; call this the *prospective value* to him of that prospect. For a given prospect, call the assignment to each person of the prospective value to him of that prospect the *value profile* of that prospect. For any strategy profile, we can thus speak of the value profile of the prospect that the strategy presents; call this the value profile *yielded* by strategy profile.

A value profile is *feasible* if it is yielded by some feasible strategy profile. In that case, it is attainable by perfect conformity to some possible social contract—namely, the social contract that tells each person to adopt the strategy assigned him by that strategy profile. The *feasible set* of value profiles is the set of feasible value profiles. For the case of two people, we can represent any value profile on paper by its Cartesian coordinates, and so we can represent the feasible set of value profiles by a set of points. A possible example is shown in the figure. A feasible value profile is *nondominated* if no other feasible value profile has a higher value on one person's goal-scale without having a lower value on someone else's.



Suppose first that set of feasible value profiles is strictly convex. The nondominated feasible value profiles then lie on the frontier of this convex set. Each, then, lies maximally in one direction—and this amounts to saying that there is a possible goal-scale on which it is maximal. (Draw a tangent to the feasible set at that point; the goal-scale is a vector that points outward perpendicular to this tangent.) Take, then, any nondominated feasible value profile, and suppose that the ideal social contract would tell each to adopt a strategy that, jointly, yields this value profile. Call this the *ideal value profile,* and call this goal-scale the *ideal goal-scale.* Among feasible value profiles, the ideal value profile comes out highest on the ideal goal-scale. (In case the feasible set is convex but not strictly, then perhaps more than one value profile will have this property.)[12]

We still might ask whether individuals can jointly reach this ideal value profile by each acting always to maximize prospects as reckoned by the ideal goal-scale. To this question the answer is no: as utilitarians have long realized, individually rational utilitarians may fail to coordinate and hence achieve an outcome that is suboptimal. (In my dissertation, my first example was a village of act-utilitarians threatened with destruction by a giant boulder; each villager rescues as many children and possessions as possible, each doing the best he can given what others are disposed to do. Jointly, though, they might have pushed the boulder harmlessly down the other side of the hill.)

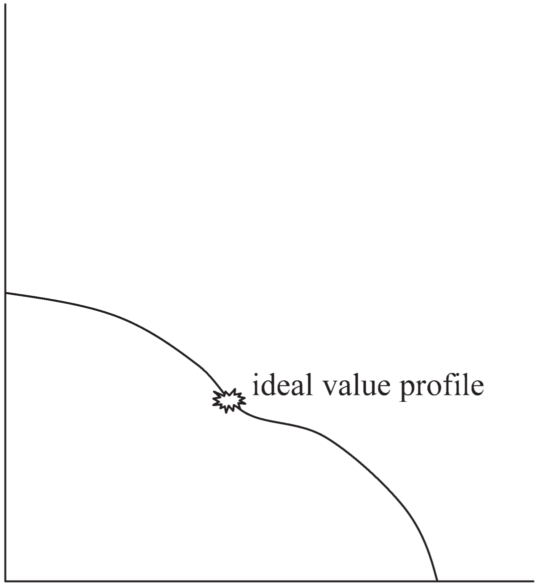
The result I appeal to, rather, is this: in abiding by the ideal social contract, each person acts always to produce prospects that are maximal on the ideal goal-scale. Or at least this is so under certain conditions, which I will sketch. This is an application of the theorem originally about utilitarianism. Take a community of perfect act-utilitarians, and suppose first that they could make binding any agreement they chose. Call an agreement that they would then make *optimal.* The theorem is that if an agreement is optimal, and if it is common knowledge that they will each keep that agreement, then each will keep the agreement even if it has not been made binding.[13] The theorem applies to people disposed always to act to maximize prospects on a common goal-scale, whether or not that goal-scale is in any sense utilitarian. The conditions are the following: (1) value from coordination only, so that no value or disvalue, as reckoned by the goal-scale, stems from anticipation, teaching, or resources being expended on

[12] See Gibbard, *Utilitarianisms and Coordination.*

[13] Gibbard, "Act-Utilitarian Agreements," 112–18; Gibbard, *Utilitarianisms and Coordination,* 186–93. The theorem was proved only for finite models.

calculation; (2) full agreement in subjective probabilities when the agreement is made; and (3) full memory as strategies are acted on.

Matters are more complex if we drop the assumption that the feasible set is convex. Then it may be that a nondominated feasible value profile is maximal among feasible value profiles on no goal-scale, as shown in the figure. If, though, any probability mixture of feasible strategy profiles is feasible, then the set of feasible value profiles will be convex.



What if parties don't all agree in their prior subjective probabilities—though each is still perfectly coherent and each counts as reasonable? The assumption that people do agree in their subjective probabilities at the time of making the agreement is crucial to the theorem about act-utilitarian agreements that I am reinterpreting, and John Broome has a result that is discouraging on this score.[14] The upshot of Broome's result in the present framework and what the consequences are for moral theory I leave for further inquiry.

Another question this Harsanyi-like result raises is what work the social coherence assumption was doing in the second Harsanyi theorem in the first place. I have argued that the ideal social contract is nondominated, and it follows from this and convexity that there is a goal-scale on which it maximizes prospects. Coherence is partly a matter of ordering, and the import of a preference ordering lies in how it constrains what's op-

[14] Broome, *Weighing Goods,* 152–54.

timal as the feasible set changes. I have considered only a fixed feasible set of prospects. We can ask, then, whether the social contracts that are ideal for different possible circumstances—the ones that, given those circumstances, no one could reasonably reject—all maximize the same goal-scale. Many contractarians will answer no.[15] If this points to an ethos of "to each according to his bargaining position," we may however conclude that it is reasonable to reject such a basis for free, unforced agreement on the basic structure of society.

This leads to difficult questions about contractarianism. Do the principles that no one could reasonably reject change as new information unfolds, information about such things as our respective social positions, needs, abilities, and the like that affect our bargaining positions? If so, it will be hard to interpret Scanlon's test: we seek principles that no one could reasonably reject, but reject at what point? If not, then we can consider a highly prospective standpoint for the acceptance or rejection of principles, and this may look a lot like Rawls's Original Position or Harsanyi's ethical standpoint. I won't, however, investigate these issues further here.

Bibliography

Barry, Brian. *Theories of Justice.* Berkeley and Los Angeles: University of California Press, 1981.

Binmore, Ken. *Playing Fair: Game Theory and the Social Contract.* Vol. 1. Cambridge: MIT Press, 1994.

Brandt, Richard. "Some Merits of One Form of Rule Utilitarianism." *University of Colorado Studies in Philosophy* 3 (1967): 39–65.

———. *A Theory of the Good and the Right.* Oxford: Clarendon Press, 1979.

———. "Toward a Credible Form of Utilitarianism." In *Morality and the Language of Conduct,* edited by Hector-Neri Casta Neda and George Nakhnikian, 107–43. Detroit: Wayne State University Press, 1963.

Broome, John. *Weighing Goods: Equality, Uncertainty and Time.* Oxford: Blackwell Publishing, 1991.

Fiske, Alan Page. "The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations." *Psychological Review* 99 (1992): 689–723.

Frankena, William K. *Ethics.* Englewood Cliffs, N.J.: Prentice-Hall, 1963.

———. "The Ethics of Respect for Persons." *Philosophical Topics* 14 (1986): 149–67.

Gauthier, David. *Morals by Agreement.* New York: Cambridge University Press, 1991.

———. "On the Refutation of Utilitarianism." In *The Limits of Utilitarianism,*

[15] Gauthier, *Morals by Agreement,* and Ken Binmore, *Playing Fair,* are examples.

edited by Harlan B. Miller and William H. Williams. Minneapolis: University of Minnesota Press, 1982.

Gibbard, Allan. "Act-Utilitarian Agreements." In *Values and Morals,* edited by Alvin Goldman and Jaegwon Kim. Dordrecht, Holland: Reidel, 1978.

———. "Disparate Goods and Rawls' Difference Principle: A Social Choice Theoretic Treatment." *Theory and Decision* 11 (1979): 267–88.

———. "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life." In *The Foundations of Social Choice Theory,* edited by J. Elster and A. Hylland, 165–93. Cambridge: Cambridge University Press, 1986.

———. "Knowing What to Do, Seeing What to Do." In *Ethical Intuitionism: Reevaluations,* edited by Philip Stratton-Lake. Oxford: Clarendon Press, 2002.

———. "Moral Feelings and Moral Concepts." In *Oxford Studies in Metaethics,* edited by Russ Schafer-Landau. Vol. 1. Oxford: Oxford University Press, 2006.

———. "Morality as Consistency in Living: Korsgaard's Kantian Lectures." *Ethics* 110 (1999): 140–64.

———. "Natural Property Rights." *Nous* 10 (1976): 77–88.

———. "Reply to Critics." *Philosophy and Phenomenological Research* 72 (May 2006).

———. *Thinking How to Live.* Cambridge: Harvard University Press, 2003.

———. *Utilitarianisms and Coordination.* New York: Garland, 1990. (Ph.D. diss., Harvard University, 1971).

———. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge: Harvard University Press, 1990; Oxford: Oxford University Press, 1990.

Griffin, James. *Well-Being: Its Meaning, Measurement, and Moral Importance.* Oxford: Oxford University Press, 1986.

Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgments." *Psychological Review* 108, no. 4 (2001): 814–34.

Hammond, Peter. "Consequentialist Foundations for Expected Utility." *Theory and Decision* 25 (1988): 25–78.

Hare, R. M. "Could Kant Have Been a Utilitarian?" *Utilitas* 5, no. 1 (May 1993): 243–64.

———. *Moral Thinking: Its Levels, Method, and Point.* Oxford: Clarendon Press, 1981.

Harsanyi, John. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (1953): 434–35.

———. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63 (1955): 309–21.

———. "Morality and the Theory of Rational Behavior." *Social Research* 44, no. 4 (1977): 623–56.

Jeske, Diane, and Richard Fumerton. "Relatives and Relativism." *Philosophical Studies* 87 (1997): 143–57.

Kahneman, Daniel, J. L. Knetch, and R. H. Thaler. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* (1990).

Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (1979): 263–91.

Kant, Immanuel. *Grundlegung der Metaphysic der Sitten* (Riga: Hartknoch, 1785). Translated as *Foundations of the Metaphysics of Morals* by Lewis White Beck (Indianapolis: Bobbs-Merrill, 1959). Standard page numbers from the Königliche Preussische Akademie der Wissenschaft edition (Berlin, 1902–1938).

Korsgaard, Christine M. *The Sources of Normativity.* Cambridge: Cambridge University Press, 1996.

Mackie, John L. *Ethics: Inventing Right and Wrong.* Harmondsworth, England: Penguin Books, 1977.

Marcus, Gary. *The Birth of the Mind.* New York: Basic Books, 2004.

McClennen, Edward F. *Rationality and Dynamic Choice.* Cambridge: Cambridge University Press, 1990.

Moore, G. E. *Principia Ethica.* Cambridge: Cambridge University Press, 1903.

Nozick, Robert. *Anarchy, State, and Utopia.* New York: Basic Books, 1974.

Ramsey, Frank Plumpton. "Truth and Probability." In *The Foundations of Mathematics and Other Logical Essays.* London: Routledge and Kegan Paul, 1931.

Rawls, John. "The Basic Structure as Subject." *American Philosophical Quarterly* 14, no. 2 (1977): 159–15.

———. *A Theory of Justice.* Cambridge: Harvard University Press, 1971.

Ross, W. D. *The Right and the Good.* Oxford: Clarendon Press, 1930.

Savage, Leonard J. *The Foundations of Statistics.* 2d ed. New York: Dover, 1972.

Scanlon, Thomas M. "The Status of Well-Being." In *The Tanner Lectures on Human Values,* edited by G. B. Peterson, 91–143. Salt Lake City: University of Utah Press, 1998.

———. *What We Owe to Each Other.* Cambridge: Harvard University Press, 1998.

Sen, Amartya K. "Rationality and Uncertainty." *Theory and Decision* 18, no. 2 (1985): 109–27.

Sidgwick, Henry. *The Methods of Ethics.* 7th ed. London: Macmillan, 1907.

Smith, Holly M. "Rawls and Utilitarianism." In *John Rawls' Theory of Social Justice,* edited by Gene Blocker and Elizabeth Smith, 346–94. Athens: Ohio University Press, 1980.

Suzumura, Kotaro. "Interpersonal Comparisons of the Extended Sympathy Type and the Possibility of Social Choice." In *Social Choice Re-examined,* edited by Kenneth J. Arrow, Amartya K. Sen, and Kotaro Suzumura, 2:202–29. New York: St. Martin's Press, 1997.

Thaler, R. "Towards a Positive Theory of Consumer Choice." *Journal of Economic Behavior and Organization* 1 (1980): 39–60.

Tversky, A., and D. Kahneman. "The Framing of Decisions and the Psychology of Choice." *Science* 211 (1981): 453–58.

Varian, Hal R. "Distributive Justice, Welfare Economics, and the Theory of Fairness." *Philosophy and Public Affairs* 4, no. 3 (1975): 223–47.