

REPLY TO COMMENTATORS

Allan Gibbard

Department of Philosophy
University of Michigan

I thank Michael Bratman, John Broome, and Frances Kamm immensely for lavishing their philosophical acumen and energies on these lectures. Each commentator raises crucial issues and presses arguments in ways that require new thinking on my part. These are the ideal people to scrutinize my attempts in these lectures, and daunting though it is to face their critiques, it is the greatest privilege a philosopher can experience to have his thoughts subject to such attention so that he can come to understand matters better.

The lectures themselves were an attempt to join two sorts of inquiry: On the one hand, I inquired into the nature of ethical judgments and of normative judgments more generally. On the other hand, I engaged in ethical inquiry proper, making ethical judgments, criticizing and refining them, and investigating their bases. One question that drove me was whether the one bears on the other. Does what we are doing when we address ethical questions bear on the answers to those questions? The basic question in ethics is what to do, I claim, and this includes what social systems to support. Can this understanding of the nature of ethical questions and thinking help us in that thinking?

I looked in particular to a longstanding debate in substantive ethical theory. We depend on the moral motivations of our fellows to foster goods in our lives and to protect us from harms. Many of the goods and harms in question are morality-independent or “nonmoral”, in that they are worth caring about apart from moral considerations. Happiness, accomplishment, and human attachments may be good examples. How much does the morality-independent good that morality can do explain the content and importance of morality? Fully, say “consequentialists” of all varieties, so that in this sense, morality is made for man and not man for morality. Strongly felt intuitions, though, seem to tell against such a view, ruling out acts that would be permissible or even required if what mattered about morality was exclusively its ties to our good apart from morality.

I asked in the lectures, then, how we should understand “intuitions” and their authority, and how this bears on what we should learn from moral intuitions. I did that by looking at two strands in moral thinking, utilitarian and contractarian. I considered the broadest questions of

social ethics: these, I say, amount to the question of what kind of social order to support from an impartial standpoint. Possibly, our moral intuitions clash with utilitarian thinking because they respond to considerations that are contractarian, to what we would have agreed on as rules to govern our dealings with each other. I examined reasons to think that we would have agreed to promote a total good that encompasses the good of each person.

The commentators raise questions about both parts of this project. On the nature of ethical judgments, Frances Kamm discusses the nature and authority of moral intuitions, and John Broome and Michael Bratman both take up aspects of my own account of the nature of ethical judgments. On substantive questions of social ethics, Kamm talks of uses of a “veil of ignorance” to approach questions of social justice, and Broome takes up my use of theorems and their import for ethics. I’ll respond to the points they raise in a different order from that of the lectures. I’ll begin with substantive issues of social ethics: the import of a veil of ignorance and of theorems like Harsanyi’s. I’ll then turn to the nature of ethical judgments and the role of intuitions. That order best allows me to conclude with the question that I find immensely difficult but didn’t much take up: Does the nature of moral questions really bear on their answers?

Veils of Ignorance

As Kamm says, I took it that “planning how to live with others should, at its base, involve planning to live” with each person “on terms of mutual respect.” I preferred, moreover, voluntary adherence to a system for achieving this aim, if we can get it. Kamm asks why, and suggests that planning to live with others in mutual respect may be in tension with conceiving morality as a means of achieving morality-independent goods.

On this question of why, I’ll only say that those were my starting points. I joined Rawls and Scanlon in proposing these as crucial aims, and I addressed readers who share the aims.¹ If these aims leave you cold, the lectures were not for you. I also joined in with Scanlon’s proposal of a more specific aim, living with others in ways one can justify to them—in ways, to put it roughly, that no one could reasonably reject. I accepted that this aim might characterize morality.

Is all this in tension with conceiving morality as a means of achieving nonmoral goods—meaning morality-independent goods, “benefits that can be appreciated in nonmoral terms”? The point of the second and third lectures was to scrutinize a powerful set of arguments that this apparent tension may, on further examination, prove illusory. We succeed in justifying a way of dealing with a person, I would think, when we show him that it gives proper heed to his good

¹ Rawls, *A Theory of Justice* (1971); Scanlon, “Contractualism and Utilitarianism” (1982) and *What We Owe to Each Other* (1998).

and that of others. As for how the person's moral good figures in such a justification, the question needs more discussion, and I'll take the question up in due course. Initially, though, note that for whether a way of treating a person is objectionable, at least some kinds of moral good seem beside the point. It won't help, say, in showing a person that we are treating his good as we owe him, to convince him that we are seeing to his virtue or good character. We can't say "Fair's fair: true, I get the money—a crass, nonmoral good—but you get the virtue." Dismissing this piece of sophistry leaves other kinds of moral good to consider, but in any case, I fully agree with Kamm when she says, "the goal of getting agreement and justifying one's conduct to others is not a mere means to securing nonmoral goods." A remaining question is how this moral aim cashes out—and that requires me to consider other things that Kamm and Scanlon say.

Reasonable rejection

Scanlon formulates contractualism as a general position, and then proceeds to develop it in his own particular direction. I myself followed Rawls in, first, adopting the general contractualism that Scanlon puts so insightfully and eloquently, and second, in specifying a way of dismissing objections as unreasonable—the "You would have agreed" retort, as I called it.² I then followed Harsanyi and expanded on him in drawing consequences from Rawls's starting points. (Some of these are consequences that Rawls himself very much rejects, but no one, I think, has found a way to make Rawls's package of theses coherent in its entirety. On this point, Scanlon and I agree.) We thus have three versions of contractualism on the table: (i) the general idea that Scanlon formulates and that Rawls and I share, (ii) the Rawls-Harsanyi version that I was advocating, which consists in Rawls's rationale carried through in the form that, I'm convinced, Harsanyi's arguments force on them, and (iii) the Scanlon version, which chiefly remains to be filled out despite his large book devoted to the project.³ I explored in the lectures whether different sources of moral concern, contractual and benevolent, converge in their import. The argument that they do requires convincing us that the Rawls-Harsanyi specification realizes the insight of contractualism—something that Scanlon and Kamm deny.

As Kamm points out, I skipped past, in the lectures, Scanlon's discussion of Harsanyi in his classic article "Contractualism and Utilitarianism" (1982). Scanlon has many reasons for rejecting the approach to social ethics that Rawls and Harsanyi shared, but to my mind the interesting one is the one in his book, his challenge to the notion of a person's good.⁴ Scanlon's earlier article was rich with illuminating insights, and Kamm reports well what Scanlon says

² Harsanyi, "Morality and the Theory of Rational Behavior" (1977).

³ Scanlon, *What We Owe* (1998).

⁴ Scanlon, *What We Owe* (1998), Chap. 2.

about Harsanyi. I don't myself, though, find the article's arguments against Rawls and Harsanyi telling, and I owe an explanation of why.

Scanlon's chief aim in that article is of course to broach his own specification of contractualism. Although, for purposes of these lectures, I accepted general contractualism as Scanlon so wonderfully formulates it, I ignored Scanlon's own way of deriving consequences. One reason is that, as Scanlon recognizes, his own way requires reaching moral conclusions prior to applying any contractualist test. The general contractual test that Scanlon and I share requires as inputs conclusions about the grounds on which it is *reasonable* to reject principles. His own approach applies piece-meal intuitions directly to reasonableness, whereas I looked to a systematic test in the spirit of Rawls and Harsanyi. Scanlon himself rejects this aspect of Rawls and Harsanyi, and thinks that no more systematic alternative will be plausible. As my agonizing in the lectures over the legitimate place of intuitions shows, I can't reject Scanlon's way of proceeding out of hand. I worry throughout the lectures over what might distinguish legitimate intuitions from dogmatic pronouncements, and that indeed is a worry I find myself with in some of Scanlon's specifications of how his version of contractualism works. If, then, we can say something more systematic about what makes a rejection reasonable, clearly that has advantages—and I was arguing that Rawls had done so.

Scanlon's treatment of Harsanyi in "Contractualism and Utilitarianism" is devoted chiefly to distinguishing the Scanlon version from the Rawls-Harsanyi version. I agree that Scanlon's own position is distinct. Scanlon notes too that many of Rawls's arguments can be given within a general contractualism—as one would expect. That leaves Scanlon's actual critique of the Rawls-Harsanyi version. In the first place, as Kamm quotes, Scanlon claims that Rawls makes a "covert transition" when he goes from the reasonable rejection test to the test of what we would have agreed to. Now I don't think that Rawls exactly hides the transition from the general idea of hypothetical agreement to his own "Original Position" as a particular interpretation, though I'll agree he is sometime obscure in explaining his motivations. The claim of a covert transition in any case can't apply to what I myself said in the lectures. I was quite overt in introducing the "You would have agreed" retort to an objection, and in delineating the circumstances where it would have force. It has force, I said, when (i) we are scrutinizing our going way of doing things, (ii) the situation in which the person would have agreed to the system is fair, and (iii) the motives for agreement wouldn't have been moral ones, but would stem from the very sorts of interests on behalf of which the person now objects.

There remains, of course, the question of whether the "You would have agreed" retort has force. I find that in my own thinking that it very much does, but if the retort leaves someone cold who genuinely understands what it involves, then I don't know anything to say that would make the person responsive. We do need to be clear how the response works: Someone objects

to proceeding by the established rules, but they turn out to be the rules he would have agreed to antecedently, before he knew whose ox would be gored, and the rules he wants applied aren't. He would have rejected the alternative he advocates, out of the very interests he now says are being short-changed. I myself find all this to be ample grounds to dismiss the complaint.

Moral goods behind the veil

Kamm and Scanlon, though, support an alternative way of glossing the “veil of ignorance” behind which moral rules might be chosen, a way that endows the parties behind the veil with moral judgments and motivations. The moral judgments the parties make behind the veil aren't, then, explained by contractualism; they serve to determine what would be chosen from behind this veil. What such a veil does is to foster impartiality in our moral judgments—and impartiality everyone in this debate agrees is needed. Also, says Kamm, “We should be concerned with the eventual position of each person outside the veil of ignorance,” and we can see the veil as “a device to get us to take seriously and compare pairwise the positions occupied by actual people beyond the veil” 19b. This amounts to requiring ideal moral judgments to be guided by full information and full and vivid realization of what is involved for each person affected. This too is a requirement that everyone in this debate accepts. The veil of ignorance test that Scanlon and Kamm advocate won't be controversial in itself, but it needs moral findings as input, and we can ask whether a Rawls-Harsanyi veil of ignorance offers a basis for the needed moral findings.

Some of Kamm's objections to the Rawls-Harsanyi test simply amount to saying that it isn't Scanlon's own use of a veil of ignorance. They are objections to keeping preordained moral conclusions out of the specification of what the hypothetical parties who stand behind the veil are like. I do not, she complains, “say anything about persons' demands that one plan to live this way with them.” If an agreement treats people disrespectfully, she says, that should rule it out 16d. People who are badly off could reasonably reject a proposed agreement 12i. People behind the veil should consider reasonable rejection 11f and the risk of being treated disrespectfully. “One could refuse, beyond the veil, to do to another what one is willing, beyond the veil, to do to oneself.” Our question, though, is when an objection is reasonable and when it isn't. These objections of Kamm's aren't relevant to a hypothetical contractarianism meant to explain the force of moral demands without assuming their validity at the outset.

It is of course legitimate for Kamm and Scanlon to argue that a Rawls-Harsanyi veil of ignorance fails to capture valid and important moral considerations. They can pertinently argue that the “you would have agreed” retort, even if true, won't show one's rejection of a principle to be unreasonable. Many of Kamm's and Scanlon's criticisms can be read this way, and so taken, they need to be considered one by one.

Kamm fixes on respect, and being treated with due respect is, I agree, a moral good that a person properly demands. Respect is a central moral ideal that both utilitarianism and contractarianism are meant to explicate, and if the Rawls-Harsanyi veil of ignorance doesn't capture it, it fails to capture a major basis of moral concern. We have to ask, though, what respect consists in. One aspect is an attitude toward a person, an emotional stance that portends constraining one's actions toward him in certain ways. Insults, undue familiarity, and the like express disrespect in this sense. The direct question for rules of conduct, though, isn't how to feel toward our fellows but how to treat them. Does hypothetical agreement from behind the Rawls-Harsanyi veil of ignorance fail to explain morally valid demands for respectful treatment?

It's bad to feel or think that one is being treated without respect, but that's a nonmoral bad, in that to see what is bad about it, we don't need to settle whether the treatment is genuinely disrespectful. Kamm's objection concerns being treated in ways that are disrespectful genuinely, and we have to ask what constitutes that. Not every action that goes against what a person wants for himself qualifies as disrespectful. Trying to get a job that someone else wants, for instance, doesn't ordinarily constitute treating him disrespectfully. What, then, makes a piece of treatment disrespectful? Kamm speaks of treating a person "as a mere thing", and others have used this phrase, but whatever it means, it can't tell against utilitarianism. Utilitarianism mandates taking each person's good into account in settling what to do, and this isn't treating the person as a thing in any usual sense. Kamm speaks too of "using others", but we use others every time we buy food or manufactured goods, and so the moral significance must attach to something more precise—like, perhaps, using people without their consent. We are left to specify what is objectionable by way of "using" people. Blowing up families, we are standardly told, is treating them respectfully if it is in pursuit of a sufficiently important goal that can't be as effectively achieved without blowing anyone up and if you foresee that you will be blowing them up as a side-effect of what you are aiming to do, rather than aiming to blow them up as a means to your goals. This may be right, but it does cry out for explanation. What makes such killing respectful?

Doesn't treating someone disrespectfully consist in riding roughshod over his legitimate moral claims, treating him in ways we owe him not to treat him? If this is right, then in order to settle what constitutes genuinely disrespectful treatment, we need first to find what we morally owe people; we have to establish what constitutes due moral consideration. That the treatment is disrespectful, then, is the conclusion of a moral assessment, not a starting point.⁵ Various moral theories tell us what we owe people, and hence what constitutes treating them with respect. Direct utilitarianism says that to treat a person with due moral consideration is to weigh his good

⁵ See Frankena, "The Ethics of Respect for Persons" (1986).

equally, along with the good of everyone else, in deciding what to do. The Rawls-Harsanyi version of contractarianism offers another answer that may be equivalent: to treat a person respectfully is to treat him in ways that he would have agreed to in fair circumstances for deciding how we are to treat each other. Perhaps neither view is right, but to establish what respectful treatment does consist in, we need to settle what we owe to each other.

Distribution and hardship

Scanlon and Kamm make other criticisms of the Rawls-Harsanyi way of filling out general contractualism, and we must ask if they have force. One might morally reject an outcome because it is terrible for some. (This criticism Scanlon shares with Rawls.) Sometimes, though, we do impose terrible hardships, as when, in a just war, we order soldiers into situations where they stand a high chance of getting killed or maimed, or when we stay out of a war even though people are being slaughtered and maimed. We justly do such things, to be sure, only when the situation is desperate—but that’s what utilitarianism would make us expect. Hardships and horrors for the few don’t often buy widespread benefits, and a utilitarianism that draws on the experience of humanity will shape its strictures accordingly. In particular, with Kamm’s own example of slavery, taking the possibility seriously that it might maximize non-moral good would involve either a blindness to what slavery is like, or a view that our snap intuitions respond correctly even to fantastic situations.⁶ Are a rational person’s reasons for avoiding slavery not urgent if they don’t include moral revulsion or a conviction that objections to it are reasonable?

Kamm says that I tend “to assimilate interpersonal to intrapersonal sacrifice” 15c and that I give insufficient heed to the “separateness of persons”. I spoke in the lectures, though, of how the term “sacrifice” already assumes that one particular arrangement is the morally privileged default. I showed how heeding the separateness of person doesn’t tell us how to make tradeoffs. Rawls famously noted that ignoring the separateness of persons might make one a utilitarian, but it doesn’t follow that heeding the separateness of person will make one a non-utilitarian. As for questions of income distribution and the like, Rawls himself adopted a mitigated maximin standard for distribution, taking it to be the proper response to the separateness of persons, but Kamm rejects that. So how are we to think morally about distribution?

A wide range of standards for evaluating distributions of income, wealth, and the like can be analyzed as maximizing the sum of some index. I’ll speak here of incomes, though whether it or something else is the best indicator of economic level is a complex question. The index will

⁶ See Hare, “What’s Wrong with Slavery”. A onetime slave himself, Hare examines a sort of case in which utilitarianism might really endorse slavery.

reflect the relative urgencies, from an ethical standpoint, of each difference in income. Suppose, for instance, that an increase of \$1000 per year for an otherwise minimum wage family is as urgent as an increase of \$100,000 per year for an otherwise median wage family. Then these differences will be represented by equal intervals on the index, and the ethical evaluation of an income distribution will go by the sum of everyone's index. The ethical view, whatever it is (within certain limits), can thus be represented by a suitably constructed index.⁷ Distribution as measured by the index will be morally indifferent, but that is just because the index was constructed to make this the case. It has little to do with the substance of the ethical view that the index represents. Even if the ethical standard comes close to maximin, an index with this feature may well exist.

The real question, then, isn't whether distribution matters, but on what scales of measurement it matters and on what scales it doesn't. For income, utilitarians, Rawlsians, and many others will all think that distribution matters, and matters greatly. They may also, though, speak in terms of a scale for measuring income by its ethical import. As measured on that scale, they can't think that distribution matters, because the scale already takes into account all ways in which they think distribution matters. As Harsanyi noted, it is nonsense to think that distribution matters as measured on a scale if one agrees that the scale already takes fully into account how distribution matters. Utilitarians propose a scale that they think has this feature.

That leaves us with the question of how to assess the ethical urgency of income differences (or differences in whatever else it is that matters ethically in economic circumstance). If Scanlon's procedure yields an answer to this, it's hard to see how. Rawls, though, does have an answer, and he may still have an answer when he is dragged kicking and screaming into the format that Harsanyi argues is forced on him. For an individual, relative urgency can be read as a matter of the gambles over income that it would be rational for him to take on his own account. Tautologically, that gives an index of urgency that will guide him rationally in self-regarding gambles, as in such things as the choice of job or career insofar as his income is what matters in the choice. When one person's income prospects trade off against another's, the question, as Harsanyi's second welfare theorem showed, is how to calibrate their scales with each other. Each person's scale indicates prospective urgency from that person's own point of view. Should we use a different scale when the prospective urgency is ethical?

Once we take on board Scanlon's critique of the notion of welfare in his book, things become less clear. Still, though, we presumably suppose that income distribution matters because income matters in a special way to the person whose income it. If, then, we can make sense of the idea of the prospects a person rationally prefers in light of this special way of mattering to

⁷ Technically, the requirement for such an index to be possible is "separability", but what this amounts to I won't go into.

him, we can again apply Harsanyi's second welfare theorem. The gauge of urgency is how urgent an individual rationally treats his income from this standpoint, and this is reflected in his "utility" scale. The prospective Pareto principle is then that prospects that are better on everyone's scale are better ethically, from the point of view of respecting each person and giving him his due.

Like things apply to saving children swamped in their canoes. (I focused perversely just on the fathers, though the kids are of course what chiefly matters in their peril. The point is that if the kids are *all* that matters, it's a no-brainer to save as many as possible. So I considered possible grounds for a father to favor his own, and that has to be something other than the value of their lives to themselves as considered impartially.) If the fathers, from the point of view where it matters which children are his, rationally find it especially urgent to have at least one child as opposed to a second or a third, then the Rawls-Harsanyi hypothetical contract procedure will reflect this as relative moral urgency. If a father doesn't himself rationally treat his having at least one child as most urgent, why should morality?

As for accepting risks and imposing risks, a reasonable system of social regulation will of course treat them differently, for reasons a utilitarian can explain. A morally sensitive person could thus, as Kamm says, "refuse, beyond the veil, to do to another what one is willing, beyond the veil, to do to oneself." Implementation of utilitarian goals will involve permissions to treat oneself in ways one couldn't treat others without their consent. It's surely not *always* wrong, though, to impose risks on others. We do so whenever we drive, or whenever we cross the street. Even with the best of intentions, lapses are inevitable, and when you cross the street, a car you failed to notice may swerve to avoid you and injure the driver. The question is what ethical standards validly govern imposing risks on others. What standards would it be unreasonable for anyone to reject? The standards consist in a kind of golden rule, I would have thought. In the case where we are all in the same boat with regard to risks and gains, why not maximize our prospects for getting what's worth wanting in life—such moral goods as respect aside. It's not disrespectful to impose the risks we would all have wanted to impose and have imposed on us in order to lead a life of amenity. Even if we care about respect as much as we care about noninjury, amenity, and the like, we'll still need a standard for what respect demands, and it would be silly to think that it demands that, out of respect for each other, we all make ourselves miserable.

The Tangent Theorem

John Broome says that the theorem I invoke in the third lecture, the "tangent theorem", isn't Harsanyi-like, and more importantly, can't do the job I ask it to do. Broome knows more about the interactions between economics and philosophy that center around these issues than anyone

else I can think of, and so I hesitate to disagree without having everything worked out in detail—which I confess I don't. But I'll sketch reasons for thinking that the tangent theorem, applied in the way I proposed, accomplishes more than Broome allows.

The argument, recall, took the form of a *reductio*. We suppose that a social contract C^* is the one that would be adopted, and that if it is adopted, each person will act to advance fundamentally different goals. Each person, that is to say, will have the policy of acting in a way that, given what he knows, maximizes the value of prospects as gauged by his own distinctive goal-scale. Since people are somewhat at odds in the goals they then pursue, the prospects that their interactions make for may be dominated, in that an alternative feasible prospect would be higher on everyone's goal-scale. In the diagram I used, there is a point on the upper-right frontier that they could jointly achieve and which does better on everyone's goal scale. Indeed there will likely be more than one such point, but choose one, and call the prospect it represents the *ideal* prospect.⁸ I then argued that they could jointly achieve this prospect by adopting a certain goal-scale in common, namely the one on which the ideal prospect is highest among the feasible prospects. Thus if implementing a proposed social contract C^* would lead different people to act with fundamentally different goals—supposing their goals coherent—then there is an alternative goal-scale that they could adopt in common, thereby each doing prospectively better by the standard of the very goals each would have if contract C^* were in force.

Broome had a number of objections to this purported finding and the significance I claimed for it. I'll take up two of them first, because I find them the most troubling, requiring careful qualifications on my part about what can be shown and what can't. First, for all the theorem tells us, even if we choose a goal-scale to have in common, it needn't be "a weighted average of individual goal scales." Second, "alterations in the feasible set will change the ideal point" and hence which goal-scale is ideal, and this creates problems that he specifies.

The ideal goal-scale

For all the tangent theorem tells us, Broome says, even if there is an ideal goal-scale that we should have in common, it needn't be a weighted average of individual goal scales. Now the argument I gave doesn't purport to establish what the ideal goal scale to have in common is like. That's why I agonized over this matter later in the lecture. Rather, the argument is a *reductio* of the claim that there is no such ideal goal scale. It starts out assuming—in order to refute the assumption—that if the ideal social contract were in force, different people would pursue

⁸ A point in the diagram may represent more than one prospect, and that is a matter that requires more analysis. A point represents all prospects that are indifferent to a given prospect on the goal-scales of everyone. But I'll speak as though each point in the diagram represents a single prospect, leaving the needed further analysis for other occasions.

fundamentally different goals, goals that can't amount to their all having the same goal-scale and applying it to different circumstances. I intended the theorem to show that this claim about the ideal social contract suffers a kind of incoherence. If the assumption were correct, this "ideal" social contract would be dominated, in the sense that there was an alternative to it that we can see must be even more ideal. For each person, that is to say, the alternative would better accomplish, in prospect, the very goals that she would have with the supposedly "ideal" social contract in force. For any prospect that isn't so dominated, moreover, there will be a goal scale with the virtue that if everyone adopted it in common and coordinated suitably, they would jointly attain that prospect. (This goal-scale will indeed be a weighted average of the goal-scales we started out with, though not every weighted average of those scales will have this virtue—and of course as Broome says, joining together to maximize on the wrong goal-scale might be worse than working at cross-purposes.)

Broome's critique makes me realize, though, that I should have been more careful. I spoke of the possibility of prisoner's dilemmas, reckoned in terms of the very goals people are pursuing. I didn't show, though, that prisoner's dilemmas definitely *would* arise if people pursue fundamentally different goals, and I couldn't have shown that. I just said that they might. People might, though, be lucky and not face prisoner's dilemmas even though the goals they pursue are fundamentally different. What I ought to have said is this: first, *if* a prisoner's dilemma arises, then people would all have done better with a common goal-scale—indeed with any of a range of goal-scales, so long as they adopted one of them in common. Second, even if no prisoner's dilemma arises, at least they wouldn't have done worse adopting any of a range of goal-scales in common. Again, doing "better" or "worse" is reckoned in terms of the very goals the person would have if the supposedly "ideal" social contract were in force. Probably too, I suspect, prisoner's dilemmas are hard to avoid when people have fundamentally distinct goals, except by fluke or in very special circumstances. I don't, however, know how to formulate this as a precise claim that could be shown true or false.

Broome mentions a different function that people might maximize to reach a given point on the frontier in the diagram. The function he gives, though, doesn't count as a goal-scale. A goal-scale treats probability mixture of indifferent prospects as indifferent. (I took it that this is a requirement of rationality, and that people will be rational in the ways they abide by the social contract, and so will have goal-scales.) The way I set the diagram up, the indifference curves that any goal-scale gives rise to must be straight and parallel.⁹ (Broome does point out, though,

⁹ The argument is this: The axes are goal-scales, and so on them, probability mixtures of indifferent prospects are indifferent. It will follow that for any possible goal-scale, indifference according to that scale will be represented by a straight line. Take any two prospects that are indifferent as gauged by goal-scale U . Then probability mixtures of them are indifferent. All probability mixtures, though, lie along a straight line in the diagram. Take, for instance, an even probability mixture of

that the function he proposes handles the problem of non-convexity, which I myself leave unresolved. Non-convexity calls for more analysis that I can yet give it, but at this point I'll just say this: If it's only when the feasible set isn't strictly convex that we shouldn't act on a common goal-scale, that in itself is a surprising finding.)

I should also speak to another question that Broome doesn't raise. Isn't the argument I have given in effect Harsanyi's own argument, invoking his second welfare theorem? I require, after all, that the common goal-scale be coherent and treat what every individual finds indifferent as indifferent. Aren't these Harsanyi's exact assumptions? Yes, I answer, but in my treatment, these features emerge as conclusions, not as assumptions. What I assume is that each individual has a different goal-scale, as a result of adhering to a particular social contract. I then say as a *conclusion* that there is a goal-scale they might have had in common, such that their having it in common would have a certain virtue. Any goal-scale with this virtue must indeed satisfy all the requirements that Harsanyi lays down for "social preference" (or if it doesn't, that's because of considerations about a variable feasible set which I'll discuss shortly.) But if it satisfies Harsanyi's conditions, that's a conclusion of the argument, not an assumption.

It may look as if the common goal-scale isn't doing much work. It only takes us to a single point in a fixed feasible set of prospects. It needn't hold steady as the feasible set changes, and so it doesn't operate as Harsanyi's "social preferences" do. In fact, though, as I am envisaging matters, the common goal-scale does considerable work. The parties who negotiate the social contract have very little information about the initial state of the world they will face. They agree to advance a fixed goal-scale as information arrives that bears on what the consequences will be. Each person, under the contract, applies the common goal-scale to many decisions taken in many states of information. As new information comes in, the prospects change, in his eyes, for how well his goals and the goals of others will be realized. Thus his prospective view of what the feasible set was keeps changing, but he goes on advancing the same agreed goal-scale. In consequence, although the map of prospects achievable by alternative social contracts looks simple and static, still the possible goal-scales that it represents each work across a vast range of informational states that, for all parties negotiating the social contract know, a person may be in at some point.

That's the reason I called the tangent theorem, in this kind of prospective application, "Harsanyi-like". To be sure, as Broome points out, the theorem is old news to any economist or applied mathematician, who thus won't find it particularly Harsanyi-like. In this application, though, it did strike me as Harsanyi-like in that, first, it is what is left when we drop Harsanyi's

prospects a and b . Its x coordinate lies halfway between a and b on goal-scale U_1 , and its y coordinate lies halfway between a and b on goal-scale U_2 , and so the point lies halfway between on the line segment joining them.

requirement that there be a social preference which is coherent, and second, it yields the result that a coherent social preference might better accomplish everyone's goals, applied as new information varies the prospective feasible set.

Varying the feasible set

Normative theorists who think in economic or game-theoretic terms differ on whether the goals to advance in common, under a justifiable social contract that avoids prisoner's dilemmas, will depend on what's feasible. Harsanyi thinks they won't, and Gauthier thinks they will. My application of the tangent theorem, Boome says, doesn't ensure that the morally ideal goal-scale for us to have in common will be independent of what's feasible. In consequence, because of the complexity of life, we can't know what the ideal goal-scale is, and might well get it wrong. If we do get it wrong, then we might do worse, as gauged by the correct ideal goal-scale, than we did each pursuing our separate goals.

I certainly agree that having a goal-scale in common is no virtue in itself, if the balance of goals it represents isn't sufficiently worth advancing. There's a general phenomenon of the "second best", that conditions that characterize an ideal may not be individually good to meet when one is away from the ideal. That's the way it is with having a common goal-scale: ideally we would have the right one, but a prospect that is top on a common goal-scale that isn't the right one needn't be better than another we attain without a common goal-scale.

As I say, the *reductio* argument that I offered doesn't tell us what the ideal goal-scale is. It just tells us that if a contractualist thinks that we shouldn't agree to a common goal-scale, his normative theory can't be right. (Any nutshell statement, of course requires many qualifications, but even those aside, this is all the argument tells us.) I of course would love to be able to establish more about what sort of goal-scale it would be ideal for us to have in common, but that's a further endeavor. Perhaps we need to look further to Kant's vision of a kingdom of ends, but I won't pursue the full vision in this reply.

We do know roughly this: that if each person has a different goal-scale and the prospective result of their interactions isn't at the frontier in the diagram, then by the standard of each of those goal-scales, adopting in common any weighted average of them that gives each of them positive weight will improve things by the standards of each person's goal scale. It doesn't follow, though, that it will be an improvement from the standpoint of justice or desirability. We didn't, after all, start out with any assumption about what is just, apart from the assumption that is shown untenable by the *reductio* for the cases where prisoner's dilemmas arise.

Will the goal-scale to advance in common depend on what's feasible? That's a complex matter. As I say, I was imagining a social contract drawn up and agreed to before any information comes along about what's feasible. Parties to the contract agree, though, in their

subjective probabilities for each way the world they will confront might turn out to be. (As Broome himself has shown, dropping this assumption stymies Harsanyi-style arguments.¹⁰) In the third lecture, though, I make no assumption that each party is looking to his prospective nonmoral good. They might, for all I was supposing, be looking to aspects of moral good. They might, as Kamm thinks they should, already have a view as to what justice requires, arrived at on grounds that don't involve what people would have agreed to if they hadn't been motivated by considerations of justice. And they might, for anything I have said, already be convinced that what's just depends on what's feasible. (David Gauthier thinks that it does, and so do adherents of the Nash solution to his bargaining problem as determining what's just.) If, then, what's feasible bears on what's just, and if the kind of hypothetical social contract that determines what's just has parties to the contract who are motivated by considerations of how an outcome relates to what was feasible, no theorem of the sort I was considering will rule this dependence out. What was feasible will then be a morally significant feature of any outcome, and this even prior to bringing contractarian considerations to bear on questions of justice. On this score, Broome is perfectly correct.

Suppose, though, as theorists like Gauthier imagine, the interests that the parties to a social contract seek to advance don't themselves involve the relation of what happens to what was feasible. Suppose that if there is a morally significant relation between the two, it has to emerge from the contractarian argument itself. Then going sufficiently prospective in our contractarian thinking allows us to consider the feasible set as fixed. What's fixed, that is to say, is the feasible set of prospects as viewed by the parties as they negotiate the social contract. In another sense, as I said, the feasible set of *prospects* varies as the agreed common goal-scale stays fixed. The parties will learn many things as they begin to lead their lives under their social contract, including things about what was feasible and what wasn't. At the outset, though, they face a single, fixed set of feasible prospects from a standpoint in advance of all social information. The variability of prospects comes only at a stage where, for whatever reasons, they have already settled on a social contract to cover every contingency, selecting a goal-scale to have in common, and they start getting information that bears on what circumstances they actually face.

The argument I gave is addressed to a contractualist like Scanlon who rejects the Rawls-Harsanyi form of contractarianism. He takes as his standard of morally justifying an action whether anyone could reasonably reject permitting it, but he rejects the moral force of the "You would have agreed" retort in the form I support. It is reasonable to reject a social order, I took it in setting up the *reductio*, if one could do better in terms of the very goals one has as a consequence of adhering to its rules and rationale, and the same is true of everyone else. In my

¹⁰ Broome's "probability agreement theorem", *Weighing Goods* (1991), 160.

discussion of the possibility that what's just depends on what's feasible, I took it that we can push the question of what could be reasonably rejected to a stage where we don't yet have information about our society in particular and each of our places in it. One is aware, though, of the possibility that things will be as they in fact turn out to be, and is rejecting or allowing rules that cover, among other things, that eventuality.

Should the argument have any force for someone who isn't even this much of a contractualist? It does amount to asking the adherent of some particular standard, "What are moral standards for?" If he thinks they are just for their own sake and that's the end of what can be said, it's hard to know what to do but walk away frustrated, or speak ad hominum to whatever moral intuitions he does have. The argument takes the form, though, that whatever is worth wanting, for each of us and with moral considerations fully taken into account, we could each better achieve it in prospect by all adopting a particular goal-scale in common. If someone is unmoved by this, I'm at a loss about how to pursue moral issues with that person—though perhaps we can find a way.

My Own Account of Normative Questions

None of the commentators are convinced by my account of what normative judgments consist in, and Bratman and Broome both focus large parts of their commentaries on misgivings over this account and objections to it.

Coherent desires

Bratman asks about wild contingencies. People act as they ought to or ought not to act in all sorts of situations, actual and hypothetical. Caesar's plight at the Rubicon was my prime example. I maintain that one's judgment on whether Caesar ought to have crossed is a contingency plan—even though one knows one will never face Caesar's plight. It is a plan for what to do if one is Caesar at the Rubicon. I maintain that contingency plans are subject to requirements of coherence. Bratman asks why this would apply to such "wild" contingency plans, to plans for circumstances one knows one won't be in. Why are even wild contingency plans subject to requirements of coherence, and desires, for instance, not?

I would deny this particular contrast; the real contrast, I say, is that the requirements on plans and on desires are different. Desires figure in a special way in planning and action, and the requirements on them stem from their special role. I desire to read quietly at home, but I desire more strongly to see the latest show, and so I go out. In this sort of way, desires weigh toward action. They have greater or lesser strengths, and the strengths of desires compose to yield one's preferences all told. The requirements of coherence that govern desires, then, are the ones that are needed for them to play this role.

Saying this requires some explanation. In the first place, the term ‘desire’ is used in various ways. “Desires” may be felt cravings, so that the feeling that one must keep an onerous promise won’t count as a “desire”. I don’t know if such distinctions among motives can be placed on a clear footing, but I have in mind a broad sense for the term ‘desire’. I count as “desires” any of the tendencies toward action that are resolved in deciding what to do, whether felt as a “beauty or a cutie,” as Ogden Nash put it, or as a stern taskmaster.¹¹ Now I don’t have firm views on the best way to construe desires in this sense, but here is one way it might go: A “desire”, we can try saying, is a decision weight. It gives a score, in effect, positive or negative, to some feature that a situation can have. This score is the “strength” of the desire. This evening I give a positive score to reading at home, and a higher score still to seeing the show. I then use these scores to tote up an expected value of each alternative open to me, and in planning, I okay any alternative with a highest prospective score and reject any that is prospectively outscored.

Now of course, our states of longing, feeling obligated, finding a prospect attractive, and the like don’t in fact come with precisely defined objects and strengths. Precise desires are ideal states, not psychic phenomena as they come to us. We need a better account than I know how to give of how an ideal role for a kind of state of mind can give rise to oughts governing it. Desires, plans, and beliefs, though, are all in the same boat in this regard. They will be vague and confused to a greater or lesser degree, whereas the story of their role will treat them as precise. The norms governing a state that have the flavor of logic rather than substance, we can now try saying, are the ones required for the state to play such a precise role. This applies, for instance, for degrees of credence as they figure in decision theory: requirements of coherence on beliefs, we could try saying, are conditions that states must satisfy to guide us to action in the way ideally characteristic of beliefs. Bayesian decision theory purports to explain the ideal guiding role of belief.

A desire is fit to play such a role, I’ll try saying, when it is precise, and it is precise when it has a definite strength and a well-defined object. The logical requirement of coherence for desires, then, is that each have a definite strength and object. For a psychic state to be a precisely delineated desire in a system of precisely delineated desires, its strength must join with the strengths of all other desires to determine preference strengths all told among ways things might be. Preference strengths all told join in turn with degrees of credence in the various ways things may turn out to be to yield prospective scores for alternative courses of action. We have desires, more or less, inasmuch as we approximate, in our choices for action, the kind of system I have just described.

¹¹ Ogden Nash wrote, “Oh, duty, duty—Why hast thou not the visage of a sweetie, or a cutie?” and David Gauthier used this as an epigram for his own theory of duty in *Morals by Agreement* (1986).

Not any possible state whatsoever that plays the kind of role I have described in moving a being would count as a desire. A robot might be set up to compute in the way I have described far more precisely than we do, but that might not settle whether the robot literally has desires. States of desiring need in addition to be like the states we know as desiring. Perhaps they need the same feel. The robot I describe will be as if it had desires, to be sure, but whether it counts as having them literally is a further question that I won't address.

Nothing in what I have been saying explains adequately how beliefs, credences, meanings, desires, preferences, and the like tie in with ideal models of them, and how this tie gives rise to logical requirements on these states. Roughly, though, the logical requirements are conditions for the states to do the job the model lays out. They are requirements for the states not to be self-frustrating. Desires, for instance, have the job of joining with degrees of credence to make for preferences and choices. Many normative requirements on desires and other such states won't have this logical flavor. There are many things it would be logically coherent to desire but crazy. Desires ought to reflect what is worth wanting in life; otherwise they are misdirected. When, though, we dispute what is worth wanting in life, the dispute gets its content against the background of the logic of desires.

Plans and wild contingencies

A complete set of precise desires would determine a contingency plan for living that would cover even wild possibilities. The plan is the one that maximizes prospective satisfaction of those desires in each possible contingency. I spoke in the lectures, though, not of the desires that generate a plan, but of the plan itself. That gave me a less complex structure to talk about, and still allowed me to find states that match okayings and beliefs in oughts. (At least there will be a match for people who are ideally rational, and so fully prone to act on their normative judgments.) The formal requirements on a contingency plan, as on a desire or a degree of belief, will be the ones needed for the state to play its role, for it not to be self-frustrating. The role of a contingency plan is to okay or nix alternatives in various contingencies, thus narrowing down one's choices should the contingency arise.

Thus we can think of structures of for thinking what to do and the like as coming in bare bones and more fleshed out versions. The bare bones version speaks simply in terms of a contingency plan. Some meat on the bones comes with a preference ordering that offer a rationale for the contingency plan: one coherently plans to do what one finds best. The full body, skin and all, comes with desires and judgments of reasons and their weights. One then coherently prefers what one finds more reason all told to want. Talk of all these levels, though, can be couched in terms of contingency plans. Judgments of reasons and their strengths, for instance, as Bratman reminds us, I treat as plans for how to weigh considerations. I think of each

layer as subject to its own requirements of coherence, and I'm inclined to think that for the most part, we can understand the requirements on each layer in terms of the point of that layer. The point of a contingency plan, for instance, as I said, is to sort out what's eligible and what isn't in order to do only what's eligible. Bringing anything special about my own view of reasons to bear on these requirements, a stratagem that Bratman proposes and then rejects, may be superfluous.

Still, I much agree that all this needs much more work, and I don't know if this layer by layer approach to vindicating requirements of coherence on judging things okay, better, or reasons can be carried through. For one thing, if nothing matters, then everything is permitted, and it doesn't matter how one arrives at one's choices. The view that nothing matters is coherent though clearly wrong, but it makes coherence superfluous. Mattering, though, pertains to reasons: some things matter in that some reasons have non-zero weight.

Why, then, to return to Bratman's main question, settle one's plans beyond anything one might need? Often there's no reason, and when there are reasons they may be various. Plans in this regard are like beliefs: many topics aren't worth forming degrees of credence on. For plans, one reason to make them that I stress is to help in setting one's standards. We can think about what matters in life by imaginatively confronting instructive situations. For the sheer logic of plans, though, what matters is not why to bother, but the things I have been discussing: what is needed for the plan to play its ideal role. Plans consist in ruling things out (where this includes ruling out ruling various things out). The requirement on a plan is that one not rule out everything, and the plan is complete just in case for each thing it covers, it either rules it out or rules out ruling it out. The direct point of ruling something out is to keep from doing it, but if one rules out everything one could do, this point can't be realized. Even in the case of a hypothetical plan for a wild contingency, the direct point of ruling something out for that contingency is to keep from doing it if the contingency arises. The occasion won't arise, one knows if the contingency is wild enough, but that's the direct point none the less. Ruling everything out that one can do in that contingency frustrates this point.

Oughts and plans: other questions

There are reasons of a specially logical kind, then, to satisfy the requirements of coherence in one's contingency planning, even for contingencies that one knows won't arise. Bratman notes a kind of circularity in saying this: Establishing this requires thinking cogently in terms of reasons, but what cogency in such thinking involves and why is the very question at issue. This kind of circularity, though, isn't peculiar to my own account of reasons judgments. It will characterize any fundamental thinking about standards of cogency. We have to be able to think already if we are to think systematically about thinking.

What is the “direction of fit” of plans and ought judgments to the world? Is it mind to world or world to mind? Both oughts and plans, in a sense, fit the world in both ways. First, both have a mind-to-world direction of fit: If a famished tiger lurks behind the door to the right, a plan to go left fits the circumstance, and so does a judgment that one ought to go left. A plan fits or fails to fit conditions; it fits whatever conditions make it the right plan, and an ought judgment likewise fits or fails to fit conditions in virtue of which one ought to do the thing in question. Both of these states of mind, then, have a mind-to-world direction of fit. Second, though, both too have a world-to-mind direction of fit: a plan’s being carried out fits the plan, and doing what one judges one ought to do fits the ought judgment. There’s a difference between these directions, to be sure: Both with ought judgments and with plans, the world-to-mind tie is fixed conceptually, and the mind-to-world tie is not. Going left fits the plan to go left, and it fits the judgment that one ought to go left. These world-to-mind ties can’t be disputed except through conceptual confusion. In contrast, though both the plan to go left and the judgment that one ought to go left fit the tiger’s being to the right, and though the tie is obvious, it isn’t conceptual. Alternative views of what the world calls for are intelligible, however crazy—for instance, thinking that the circumstance calls for getting oneself eaten.

It’s another kind of “mind-to-world” tie, though, that philosophers might have in mind for ought judgments. We can gloss the “world” as including what one ought to do, and the judgment that one ought to go left can then fit the “fact” that one ought to go left—and the tie is conceptual. I can’t object to this: It’s a feature of the “world” that one ought to go to the left, a deflationary schema guarantees, just in case one ought to go to the left. And that one ought to go to the left isn’t made so by one’s mind. This contrasts with the plan fragment, “Let me go to the left!” where talk of a feature of the world clearly isn’t in order and we can’t properly speak of “the fact that let me go to the left”. A remaining question, though, is whether this contrast is deep or a matter of grammatical form. Indicative forms embed freely and imperatives don’t, and facts in the broad, deflationary sense are the shadows, as it were, of this grammatical form. Once we put “Let me go to the left” in indicative form—say, as “I am to go to the left”—we can say corresponding things about oughts and about plans. The plan to go left fits the “fact” that one is to go left, and the tie is conceptual.

Next, on normative disagreement: In *Wise Choices* I stressed interpersonal coordination. But though crucial to ethics, coordination may matter far less for normativity in general. Disagreement is the key, as Bratman says, and in the interpersonal case, disagreement must be understood as part of jointly thinking together, putting our heads together on how to live. Bratman gets my views on this just right. He then asks about impasses: if we face one, can we still think we are disagreeing? I find this puzzling, and I have spend parts of both my books on it. I think I want to say this: If it were clear where the impasses lie, would there then be point to regarding us as coming up with separate answers to the same question, a question of what

matters in life and so of what to do if one is you or if one is me? I could imaginatively debate the issue in my mind and imagine your voice as part of the debate on the question I am pondering. But I regard myself not really disagreeing with you, but as disagreeing with the side of me represented by your voice.

Weakness of will

On this topic, I am uncertain what is the best thing to say. One approach I find clearly unsatisfactory: to say that ruling out an action in the course of planning is one thing, and thinking one ought not to do it is another. To say this leaves it a mystery why not to scrap all ought thoughts as having no clear content. It also means we could have two parallel sets of concepts, the plan-laden concepts that I describe and show, I think, to be possible, and then these mysterious but distinct ought-laden concepts. Why have both? Because of weakness of will, goes the objection to my account. Weakness of will is irrational, though, and so if we have ought-laden concepts as well as the plan-laden concepts that would suffice for all practical purposes, this can only be in order to give us an extra way to be irrational: form the conclusion that one “ought” to do such-and-such, and then don’t do it.¹²

Don’t cases of weakness of will, though, show that, for better or worse, we *do* have these distinct, normative concepts? They may not make any sense, but don’t we none the less have them? Well, I’m not sure. People insist that, contrary to what Socrates thought, they do things at the very instant of being firmly convinced that they ought not to do it. They may, at the instant of action, change their minds about what to do at that instant, but they don’t, they insist, change their minds about what they ought to be doing. Bratman speaks of having a second glass of wine while he thinks that he shouldn’t. Now of course I do have to recognize this phenomenon, but what are we to make of it? Is he really having a thought with clear content? Perhaps his “ought” is not the general ought that I’m trying to explain, but one that takes into account a restricted range of considerations.

In my 1990 book *Wise Choices, Apt Feelings*, I worked to accommodate what such an objector maintains.¹³ In the terms I’m now using, the approach amounts to this. Ought judgments are planning judgments, but not of the whole mind but a part of it, a part I called the “normative control system”. It’s the part of the mind that makes contingency plans. But when it comes time to act on a contingency plan one has, other motivations come to bear: fears, cravings, feelings of embarrassment or shyness, and the like. You think you ought to forgo the second

¹² Scanlon thinks there are uses for a distinct ought concept; see his “Metaphysics and Morals” (2003) and “Reasons and Decisions” (2006) with my “Reply to Critics” (2006).

¹³ In my treatment of weakness of will in the book, I was responding to challenges Bratman pressed on me in a wonderful series of lunchtime conversations we had while I was first writing the book.

glass, in that the planning side of you mandates forgoing it. This is the side of you that both looks to a situation regardless of whether you are now in it and motivates you when you are in that situation. Appetites, social yearnings, and the like, though, work on motivations right now in a way that they don't work on plans. Planning for a situation like your own right now, yearnings work on you but you say to yourself, "Sure, drinking more would be convivial, but in the morning I'll feel horrible. So when the time comes, let me forgo the second glass!" I still accept all this when the time does come. I accept, in effect, "When the time comes, let me forgo the second glass!" and I accept "The time has now come." The side of me that reasons what to do in situations concludes, "Let me forgo the second glass!" But appetite and yearning for conviviality work on me, and the totality of my motivation doesn't sufficiently heed the mandate of my planning side.

When I wrote *Thinking How to Live* (2003), I worried about whether it was psychologically realistic to think that there is a distinctive normative control system. As I might now say, my worry was whether there is a distinctive plan-responsive aspect to motivation, as opposed to responsiveness to craving, yearning, fear, embarrassment, and the like as they work on action but differently on planning for action. It seemed to me also that failure to think in a unified, coherent way is ubiquitous in our experience, and explains why we would experience some situations as showing weakness of will. It's not that there's some clear judgment of "ought" that we make which then fails to prevail in our motivations. Even if I'm yelling to myself as I start on the second glass, "I ought not to do this!" there's not something I mean apart from the injunction "Don't do it!" The timorous Penzance policemen sing, "Yes, forward on the foe!" even though, as Major General Stanley observes, they don't go. In a way, they accept what they are saying, and in a way they don't.

I'm not sure which is the better way to handle situations of "weakness of will". More recent evidence may support the line I took in *Wise Choices*. The evidence for "dual process theories" supports a psychologically real distinction between will power and other motivations.¹⁴ Some might maintain that I could have my will steadily directed toward a policy, in my hypothetical thinking on what to do, and still think that I ought to do something else. I'll agree that there may well be senses of 'ought' for which such a thing is possible—a specifically moral sense, for instance. But I think there's also a "flavorless" sense of the term 'ought', a sense in which what I "ought" to do is what it makes most sense to do, everything considered. This is the sense, I think, that Ewing identified. For this sense, I can't make sense of someone's genuinely believing that he ought to do a thing while steadily willing to do something else. If someone claims such an opinion, he either doesn't have this sense in mind or is oblivious to his real convictions.

¹⁴ I thank Chandra Sripada for calling my attention to these developments; see his "Weakness of Will and the Divided Mind" (submitted). Howard Nye has also urged that I should stick to the *Wise Choices* account of weakness of will.

Identifying the attitude

John Broome has a somewhat different objection to my account of normative concepts. He thinks that my argument for the very possibility of the kind of concepts I describe fails, that I haven't proved that the planning states of my theory exist—except by helping myself to familiar normative concepts like OUGHT. In particular, he says, I haven't identified the state of mind of “okaying” an act, except as believing the act to be okay. He denies the independent intelligibility of thinking, hypothetically, what to do in a wildly hypothetical situation and okaying some alternatives while rejecting others.

Such hypothetical planning, though, it seems to me, is not hard to grasp. Suppose, fantastically, to use my stock example, you are forthwith to be Julius Caesar at the Rubicon, and now, in this frame of mind, think what to do. I don't find such thinking hard to understand. Rejecting some alternatives and ruling out rejecting others might well be stages toward hypothetically picking a course of action. (Indeed, it's hard to think why the subsequent stage of hypothetical picking, forming a full hypothetical intention to do one of the things one rejects ruling out, might ever be worth bothering with.) “Okaying” an alternative, in this hypothetical frame of mind, is just rejecting ruling it out by preference. Indeed in the case of action, if we can understand preferences, we can understand okaying and rejecting: to reject crossing the Rubicon, for the hypothetical case of being Caesar, is to prefer being Caesar at the Rubicon and holding back to being Caesar at the Rubicon and crossing. Such hypothetical okaying or rejecting may of course be idle—but it needn't be. It may amount to rehearsal for kinds of decisions one might have to make, refining one's powers of thinking what to do.

Suppose, though, Broome were right that attitudes like okaying can only be identified in the first place as beliefs. It follows, he thinks, that “if they are rational, they cannot help having the structure of rational beliefs anyway. Attitudes that are identified by their cognitive aspect cannot, if they are rational, help standing in the relations that rational cognitive attitudes stand in. The explanation of why they stand in these relations is that they are rational cognitive attitudes.”^{6f} But this, I say, is no explanation at all. True enough, if someone becomes a murderer, he kills a person. Identified as becoming a murderer, we might say, he can't help but be killing. This doesn't much aid us, though, in understanding murder. Likewise, it's true enough that if something is a belief, then it has the features of belief. But that leaves everything to be explained—including how there can be beliefs with the “queer” features that Sidgwick, Moore, and others identified.

If I am right about how normative beliefs work, then to be sure, we should be able to identify ought beliefs as Broome advocates, just as by their subject matter, and to speak of aspects of the “world” that they are about. Almost trivially, dog beliefs concern doggy aspects of the world, and ought beliefs, if they are in good order, concern oughty aspects of the world. But how can

there be beliefs with the features that Sidgwick, Moore, and others identified in normative beliefs? Are they perhaps just pseudo-beliefs, like beliefs about gremlins?

We might have thought we needed ought beliefs to figure out what to do. On an approach that identifies them as beliefs and leaves it at that, however, they don't seem needed. I can ask myself what to do, settle on reading the newspaper, and my belief that I ought to be working on a reply to Broome need have nothing to do with it—according to many philosophers. I reject any plan to hit my thumb with a hammer, and to do this, why would I need to believe, even implicitly, that I “ought” not to do so or that it would be a “bad thing” to do so? I just need the belief that it would hurt like hell, along with the absence of any countervailing beliefs (like that it would keep me from getting sent to a war in which I was likely to be killed or maimed). It's true that if I do what I think I “ought” not to do, I'm then “akratic” and thus “irrational”. But that's no more than to say that I'm doing something I think I ought not to do, that lacks a certain feature. How does this differ from picking a car that I think lacks a certain feature, such as having a gremlin? Whence the special significance of the ought feature?

It's true that if you, like any of us, have ought beliefs, then you regard oughts as important, but this needs explaining. (Some philosophers think there is such a thing as an irrationalist who has ought beliefs but doesn't care; I myself think that any halfway plausible candidate for being such an irrationalist is just mixed up in his use of the term ‘ought’, and doesn't know what he's talking about.¹⁵) Thus if someone asks questions about oughts that cry out to be asked, I don't find the answers that someone with this approach can give satisfactory.

I have argued, then, both that we can identify the attitude of okaying in an informative way, and that we are philosophically in a bad way if we can't. Is the mental attitude of okaying non-cognitive? In my 1990 book *Wise Choices, Apt Feelings*, I did use that label for my theory of normative terms, but after that, I became increasingly puzzled about what the term was supposed to mean. (One eminent psychologist said, after some thought, “I guess when I use the word ‘cognitive’, I mean it's complex.”) As for normative “facts”, in the ordinary sense of the term, there aren't any: When the detective admonishes “Just the facts, ma'am,” it isn't responsive to say, “The creep deserved it, and that's a fact!” In a philosopher's deflationary sense of the term, though, there are indeed normative “facts”, if I'm right: “That pleasure is good is a fact” means, in this philosophers' sense, just “Pleasure is good.” It's quite right, though, that in my explanations, I don't start out assuming that there are normative facts, even in a deflationary sense of the term ‘fact’.

I do agree with Broome that there are two distinct, separately intelligible questions: “What shall I do?” and “What ought I to do?” which call for two distinct, separately intelligible sorts of

¹⁵ See Lenman, “The Externalist and the Amoralist” (1999) and my *Thinking How to Live* (2003), p. 12.

answers: an intention and an ought belief. Here the question “What shall I do?” shouldn’t, of course, be read as calling for a prediction; it calls for picking an alternative. Now forming an intention or picking a course of action can, on my view, come in two stages: rejecting or “okaying” various alternatives, and then if one okays more than one, picking among the alternatives that one okays. The second, we can say, is forming an intention, and the first stage, on my view, pertains to ought beliefs. One needn’t think that one ought to do what one intends; one may just think it okay to do and pick it, thinking one or more alternatives okay to do too.

Intuitions

I myself end up relying on intuitions, cautiously and critically, but I have two main sorts of initial worries. One is that even our strong intuitions turn out to be inconsistent. It’s hard to see which intuitions, if any, can emerge undiscredited from the inconsistencies we discover. The second is John Mackie’s worries over “queerness” and superfluity: why think the universe contains the strange kinds of facts that we seem to intuit, when the psychology of seeming intuitions doesn’t require their veridicality to explain our having the convictions we do and the strength of those convictions.

Veridicality judgments as plans

My answer to the queerness worry is that the veridicality of intuitions, their de jure genuineness, is a planning question. It’s a question of which of our convictions to rely on. Though it may be legitimate to speak of “normative facts that obtain independent of us”, that will only be in the end and with a proper understanding. It isn’t the place to start, I say, in explaining the psychology of our seeming normative intuitions. To start with “normative facts” invites the queerness and superfluity challenges, and leaves us with no cogent answer to these challenges. It’s quite otherwise when we see that a de facto intuition’s de jure status is a planning issue. You or I come to a view on these matters when we come to plan to rely on certain sorts of de facto intuitions. Settling on relying on a judgment isn’t coming to have a psychological belief about oneself; it is coming to adopt a plan. I myself have plans along these lines, even if they are scattered and ill-formed, and I seek to put before readers and listeners considerations capable of persuading them to be comfortable with having such plans—but not too comfortable.

Frances Kamm, in a crucial way, doesn’t quite get right the way I try to work all this out. “Suppose,” she says, “there were no sense in which intuitions are genuine other than that we would decide, when we are in a state of full information, alert, dispassionate, etc., to plan to rely on them?” 5c. I agree with her that the consequences would be untenable, but I myself suppose no such thing. The sense in which some de facto intuitions are genuinely de jure, I say, is that they are intuitions to rely on. That a seeming intuition is one “to rely on” is different from the

psychological claim that one would have it in such-and-such circumstances, or that in such-and-such circumstances one would decide to rely on it. Claims that an intuition is genuine are part of planning, not of coming to beliefs about the psychology of planning.

I'm not sure whether my views on ethical intuitions should "give comfort to the many who ordinarily place stock in intuition". One picture is wrong, I claim, and if it guides these many, they should rethink. It's not, I say, that there's a unified, coherent way of thinking about ethical questions to which we confusedly respond and which we can bring to light by a method of intuitively supported hypothesis and intuitive counterexample. Our intuitive responses do often have rationales, but there is no single unified way those rationales fit together—no way that explains the shape of our responses. For one thing, our responses are inconsistent. The same can be said, to be sure, for our sensory responses, as with the Mueller-Laier illusion where the apparent lengths are shown false by measurement with a ruler. In the visual case, though, we can form a consistent view of the objective world revealed by rulers and the like, and this objective world enters into the explanation even of the illusion. (Quantum findings may not fit this pattern, but I'll pass over them in silence, since I don't know what to make of them.) We could try telling a story of ethical intuitions with some normative way things objectively are playing the role that geometric layout plays for vision, but we have too many candidates for what the objective normative world might be. Perhaps, as Sidgwick and Hare thought, it is hedonistic universal act-consequentialism. Perhaps it is some deontological pattern, as current philosophical opinion would have it. We have to settle what's veridical and what's distorted in our responses to the normative facts, and a psychological account of the workings of normative intuitions won't by itself yield an answer. We are learning more and more these days about how ethical intuitions work psychologically. They stem, it now appears, from a clash between at least two sorts of psychic systems, one utilitarian and one deontological—or that may be a good first approximation.¹⁶ Which system prevails in a given case depends not on some standard that might be a plausible candidate for the objective truth of how considerations weigh against each other, but on such things as how close to the person one kills one is standing. Moral intuitions are not, in their psychology, responses—even distorted responses—to an ideal pattern.

I think, then, that we are stuck with a choice between intuitions as sheer psychic happenings with no status as information givers, and intuitions as I picture them. My own view ties in closely with the arguments of Sidgwick, Ross, and others for the indispensability of intuition in ethical thinking. Since we can't regard de facto intuitions as causal responses to their truth-

¹⁶ Greene *et al* (2004), "The Neural Bases of Cognitive Conflict and Control in Moral Judgment".

makers, we should fix on their role in thinking how to live. Whether a de facto intuition is an intuition de jure is a question, we should realize, of what sorts of judgments to trust.

Assessing intuitions

When it comes to killing and letting die, certainly our moral reactions to the two are different. That leaves us with the question of how to act in light of the contrast. We can ask ourselves how to treat the difference in reactions, and in particular, whether to take it as a fundamental guide to action. I don't entirely know what to think on this score, but the following thought experiment seems to me to be highly relevant. Imagine we somehow erased our special horror of killing. Would we lose anything that we can understand independently of the special horror of killing, of our finding killing as such horrifying? We feel it's worse to be killed by someone than not to be saved by them, but experience equal, I find it hard to take this intuition seriously. But even so, as we all know, social prohibitions on killing are often highly effective but all too often not. Conditions where they are not are horrific in terms of the scale of deaths of people in their prime and the fear in which people live. Perhaps we can relax the prohibitions in special cases, but we'd better be careful not to undermine the special feelings of horror that protect us.

I agree with Kamm that an intuitive judgment "is no less objectively true if awareness of the factors and reasoning that justify it comes after the judgment than if such awareness comes before." I don't, however, think that this goes "contrary to what Haidt suggests" 2b. My worries and Haidt's aren't that judgment comes before awareness of reasoning: we fully allow the possibility that Kamm points out. Haidt and his co-workers, though, find that, for instance, people cling to their intuitive condemnation of incest even when they are shown to their own satisfaction that all the grounds they thought they had were bogus. The subjects feel "dumbfounded". Still, perhaps they should think that there remain non-bogus grounds that they haven't been able to discern.

Kamm stresses that, as I would put it, de facto intuitions should spur us to look for a deep rationale that vindicates them. With this I thoroughly agree: There's a strong probability that a de facto intuition ties in with something well worth caring about. Once we identify a candidate rationale, we can think whether to live in accordance with it when it conflicts with other candidate rationales. We may, in some cases, rightly conclude that the rationale gets at something with an important bearing on how to treat each other.

Kamm asks, on my view, "What reason is there to think that one plan about what to do and feel in the realm of morality would be any better than another plan?" Well, first note that any answer to this must either depend on more basic claims about reasons and what's better than what, or have some independent plausibility, some plausibility that it doesn't get from a further claim. What's wrong with a plan to touch a hot stove? That I'd be burnt and it would intensely

hurt. That's a reason. What makes it a reason, though? Why shun anguish? These aren't questions with further answers. To think this, I say, is to weigh anguish heavily against any course of action. Don't you agree with this weighing? What more is there to ask? To think this *de facto* intuition an intuition *de jure* is to trust such planning. Don't you agree with me to trust it? So what reason is there to plan not to touch a hot stove? The obvious one, that it would hurt intensely.

That's uncontroversial, I hope, and my own point was about what we are claiming when we say this. No direct gloss would be informative, but I can say what sort of state of mind this judgment about reasons is. It consists in planning to weigh anguish against a course of action. Its basicness consists in its not being rooted in something further to be done or to be sought.

Could I be wrong that it is bad to touch a hot stove? I am fixing on the most unproblematic aspect of how to live, and so for this particular judgment, I don't see how I could be wrong. On many matters, though, I might certainly be wrong. I might be wrong that human goods (and the goods of other sentient beings) underlie the valid demands of morality, and that the goods in question can be appreciated aside from being already committed to morality. I might be wrong that appreciating wonderful poetry is better than an equal, drug-induced appreciation of push-pin. What does being wrong on moral matters consist in? Truistically, it is believing what is not the case. It is, for instance, believing that with pleasure equal, push-pin is as good as poetry, when pleasure equal, push-pin is not as good as poetry. Beyond this truism, I can't say anything direct and utterly general. I hope, though, that I have said what it is to *believe* a judgment wrong. As for which judgments *are* wrong and how to tell, those are questions of how to live and how to think about how to live. Those are the kinds of questions I was addressing in the second and third lectures.

Metaethics and Ethics

That brings me to a question that the commentators don't address and that I find extremely difficult. Does the nature of thinking ethically bear on the content of ethics? Should understanding what ethical thinking consists in make any difference to our ethical conclusions? That's not a question we could answer on the basis of metatheory alone. My metaethical claims don't entail directly any normative conclusions. On the other hand, we can't antecedently rule out that the judgments we make will respond to our view of what we are doing, and that this responsiveness is proper.

It seems to us clear we shouldn't push a person in front of a trolley even to save five people with certainty. We know now that making this judgment is a result of emotional centers in the brain overpowering centers that operate in a more or less utilitarian way, and that these emotional centers are highly responsive to such matters as how close one is to the person one

kills. Also, in ways that haven't been studied neurologically so far as I know, the brain delivers a firm judgment that sheer literal nearness isn't morally relevant. Everyone agrees that a strong emotional revulsion to close-range killing is a good thing for us to have—even if that's only because it works, mostly, to correct for such things as wishful thinking and misjudgments of evidence that can distort utilitarian calculations. Our question concerns our revulsion-infused judgment tendencies to rule out this instance of killing to save. Are they intimations of wrongness even if for no further reason? Or are they useful emergency danger signals, good for the most part in keeping us on the right path, but sending the wrong message in this particular instance?

That's one question, but now I'm asking a further question *about* this one. Should it make any difference which of the following two our question is? (i) whether the emotional response is an indication of wrongness in the way, say, that arithmetic judgments respond to how things are with numbers, or (ii) whether to give the response fundamental weight in our thinking how to deal with each other. Nothing rules out conceptually an answer of either yes or no to this question of whether the nature of intuitions bears on how to assess them. As a matter of sheer conceptual requirements, anything at all might bear on what to do and what to weigh towards doing things. Still, once we put our ethical questions as ones of what to do and how to feel about things we can do, we may take up substantive questions of ethics in a different frame of mind.

We feel the wrongness of killing more strongly than we feel the wrongness of letting a person die, even when that clearly is the only difference that could matter. This difference in our responses is probably a good thing, and if it indicates the special wrongness of killing whether or not we can find some further ground for abhorrence, we'd better take heed. If, though, the question is what to do and why, we may find the fundamental import of the kill/let-die distinction more suspect. True, we'll have to take some things as basic grounds for action, but why this? Someone is just as dead in either case; indeed we have stipulated that there is no reason to treat the two actions differently *except* whatever it is that makes one a case of killing and the other a case of letting die. First, then, why care if you are the one who will in either case be dead? And second, if there's no good answer to this, why care if you are the one who must choose to kill one or let the other die? That being alive and the things it allows matter is likewise an intuition, true enough, but it isn't one that melts away if the question becomes what to want and why, what considerations to weigh for and against actions, and what responses to treat as guides to action. Intuitions treated as visions of how things stand morally, in contrast, aren't as open to the challenge "Why on earth heed that?" as are seeming answers to what to want and why. Seeing the question as how to live and the grounds for answers as ordinary facts may make us more attuned to what really matters in the ways we treat each other.