Comments on Allan Gibbard's Tanner Lectures

JOHN BROOME

It was a great privilege to be invited to Allan Gibbard's lectures, and to comment on them. Gibbard has made extraordinary contributions to widely separated areas of ethics. He is a leading figure in the expressivist approach to metaethics, and he has also made huge advances in first-order ethical theory, often by bringing to bear the formal methods of decision theory and game theory. His Tanner Lectures integrate these apparently very different subjects. They show us how his first-order theory is motivated by his metaethics—specifically by his view that ethical questions are planning questions.

I admire this integration, but I am sorry to say I have not been able to imitate it. I have decided to take up two separate issues, one from Lecture I and one from Lecture III.

Comment on Lecture I

Take an ought sentence such as 'Brutus ought not to have conspired against Caesar' or 'I ought to be careful here.' Philosophers used to worry about whether sentences like this could be true or false, and many denied they could be. But these days we worry less about that. Most philosophers nowadays think it is not so hard for sentences of a particular class to meet the criteria that allow them to count as true or false. They need only to participate in our thinking

and talking in characteristic ways. For example, we need to treat them as subject to truth-functional logic. We need to be able to make sense of disagreement about them, when some of us assert a sentence and others deny it. We need to recognize that a sentence might be true even though no one is in a position to assert it justifiably. And so on. Since ought sentences meet these standards, they can be true or false. Consequently, we can have cognitive attitudes toward these sentences or toward their contents. We can believe or disbelieve them, or what they say.

That is no longer very controversial. We can accept that ought sentences are true or false because of what we do with them. But this leaves us with the task of explaining why we do those things with them. How come these sentences participate in our thinking and talking in ways that are characteristic of truth?

If we were dealing with sentences about natural things, the answer to this question would emerge from the relation between these sentences and the facts of the world. Sentences about natural things are true or false in virtue of their relation to the world. That explains why our thinking and talking treats them in the ways that are characteristic of truth. We could give a parallel answer for ought sentences: We could say they are true or false in virtue of their relation to the normative facts of the world. But Gibbard and many other philosophers find that answer fantastic. It seems incredible to them that the world contains such normative facts. So they look for an alternative explanation.

Gibbard offers one. He offers an explanation of why we use ought sentences, and why we use them in such a way that they earn the right to count as true or false. His detailed explanation appears in his book *Thinking How to Live* (Harvard, 2003). His Lecture I contains an outline of his explanation.

It is that these sentences help us to plan our lives in general, and to plan what to do on particular occasions. He says that, when we utter an ought sentence, we are (to a first approximation) expressing a partial plan. It is a natural fact about us that we make plans.

Consequently, as Gibbard emphasizes, his explanation of the truth of ought sentences does not assume that anything exists apart from the natural world. It does not assume there are normative facts in the world.

Gibbard explains in detail how, as our thinking and talking use these sentences to express our plans and develop our planning, they endow the sentences with the characteristics of truth and falsity. For one thing, in *Thinking How to Live* he provides a semantic theory that explains how they participate in truth-functional logic. The details do not matter here.

Since ought sentences have the characteristics of truth and falsity, that explains how we have the attitudes of belief and disbelief toward them. These cognitive attitudes are explained on the basis of the noncognitive attitudes that are involved in planning. Indeed, they simply *are* those noncognitive attitudes in another guise. To believe you ought not to feel resentment is just to plan not to feel resentment. To believe Brutus ought not to have conspired against Caesar is just to plan not to conspire against Caesar in the counterfactual situation of being Brutus as the Ides of March approach.

Gibbard argues that planning attitudes, if they are rational, are connected together in a structure that mimics the structure of rational believing attitudes. This explains why the sentences that express the planning attitudes have the logical structure of truth. It allows us to treat our planning attitudes as ought beliefs. These attitudes are fundamentally noncognitive, but they earn the right to count as cognitive. Each is a planning attitude and also a believing attitude. That is to say, each attitude has the property of being a noncognitive planning attitude, and also the property of being a cognitive believing attitude. But the property of being a noncognitive attitude is the fundamental one, in that it explains the property of being a cognitive attitude.

So ought beliefs are plans, to a first approximation. There are two reasons that this is only an approximation. One of them is mentioned by Gibbard himself. Not all plans are ought beliefs. When

you are indifferent between two options, you will not believe you ought to take one of them, nor that you ought to take the other. But unless you are as foolish as Buridan's ass, you will plan to take one or else plan to take the other. So you may plan to do something without believing you ought to do it.

Because of this, Gibbard's theory is not exactly as I have described it so far. Gibbard does not say that ought beliefs are exactly attitudes of planning. Instead, he says they are founded on what he calls a 'valenced' attitude. This attitude might be called 'okaying.' (This is my name; Gibbard does not use 'okay' as a verb.) Okaying is a noncognitive attitude. It is like the attitude of planning, but weaker. Buridan's ass okayed eating the left bale and also okayed eating the right bale, but it did not plan to eat the left bale and also plan to eat the right one. Indeed, it did not plan to eat either. A more sensible creature would have planned to eat one or the other, but it would not have planned to eat one and also planned to eat the other.

The noncognitive attitude of okaying corresponds to the cognitive attitude of believing okay. I hesitate to attribute to an ass cognitive attitudes that have a normative content, because it may not be clever enough to possess them. But an adult human being can possess them. So when an adult human being okays an action, she believes it is okay. We can now also specify the cognitive attitude of believing one ought to do something. To believe one ought to do something is to okay doing it and decline to okay not doing it. Once again, it is the noncognitive attitudes of okaying and declining to okay that are fundamental. But because they participate in our thinking and talking in ways that are characteristic of truth, they earn the right to count as cognitive attitudes too.

That is one reason that Gibbard's initial formulation is just an approximation. The second is one he does not mention. The attitude of believing you ought to do something simply is not the attitude of planning to do it. Gibbard says that thinking what you ought to do is thinking what to do. But it is not. Thinking what you ought to do is to ask yourself, "What ought I to do?" whereas to think

what to do is to ask, "What shall I do?" These are different questions, and they may have different answers.

"What shall I do?" is not a typical sort of question. To answer a typical question, you simply express a belief. If you ask yourself, "Where are my keys?" and answer, "Beside the phone," your answer simply expresses your belief that your keys are beside the phone. But if you ask, "What shall I do?" and answer, "Read the newspaper," your answer expresses more than a belief. It expresses an intention to read the newspaper. Gibbard would call your question a 'planning question.' In asking it, you call on yourself to form an intention. True, by the time you come to answer, "Read the newspaper," you believe you will read the newspaper. That is because, in forming your intention to read the newspaper, you acquire the belief that you will do so. Your answer expresses your belief as well as your intention. But the special feature of your question is that it calls on you to form an intention.

"What ought I to do?" is a typical question, at least on the face of it. It calls for an answer that expresses a belief. If you answer it, "Start writing that lecture," you express the belief that you ought to start writing that lecture. However, Gibbard treats this too as a planning question. Maybe it is, but it is certainly not the same as the simple planning question, "What shall I do?" If its answer constitutes a plan, it does not constitute a plan of the simplest sort, such as your plan to read the newspaper. At the same time as you answer the question "What ought I to do?" with "Start writing that lecture," you might answer the question "What shall I do?" with "Read the newspaper." Then you would plan to read the newspaper while believing you ought to start writing that lecture.

If you give these answers, what you plan to do is something other than what you believe you ought to do. This means you are akratic. We have to recognize that akrasia is possible. It follows that thinking what you ought to do is not thinking what to do.

This does not mean Gibbard is wrong to treat our ought beliefs as fundamentally noncognitive attitudes. They may be, but they are

further from ordinary planning attitudes than he suggests. Gibbard only says they are "like" planning attitudes. That remains possible. They could resemble planning attitudes, even though they are rather far removed from our ordinary planning attitudes. They could be some sort of *ideal* planning attitudes—not what we actually mundanely plan but what we plan ideally in some way or other.

Suppose these attitudes are some sort of ideal plans. Gibbard recognizes anyway that they must be idealized, because of another feature of them. We have beliefs about what all sorts of people ought to do in all sorts of circumstances. If these are to be construed as plans, they must be plans that are conditional on remote and impossible conditions. For instance, if you believe Brutus ought not to have conspired against Caesar, you must have a plan that is conditional on your being Brutus. As a planning attitude, this is very idealized.

I think the remoteness of these ideal attitudes from ordinary planning raises a serious problem for Gibbard. These attitudes are the foundation of his account of normativity. They are fundamentally noncognitive, though they can earn the right to count as beliefs. But how can we identify these attitudes and know what attitudes they are?

We can generally identify people's ordinary plans rather easily. If you plan to read the newspaper, your plan is a disposition of yours that will, among other things, typically cause you to read the newspaper. That makes it easy to recognize. The disposition that constitutes a plan is complex, but its details can be spelled out; many of them are spelled out in Michael Bratman's *Intention, Plans and Practical Reason* (Harvard, 1987). In any case, we are very familiar with ordinary plans as part of the regular commerce of our lives. Often, we easily recognize our own plans and other people's, through our ordinary understanding of our psychology and theirs.

But frequently, the ideal attitudes Gibbard calls on can be identified only through the properties they have as cognitive attitudes. Take the attitude of okaying something. This is the noncognitive

attitude that founds, explains, and constitutes the cognitive attitude of believing the thing is okay. Are you familiar with this attitude of okaying? I think you will be able to recognize it only as the cognitive attitude of believing the thing is okay. I do not think you have any other way to grasp what this attitude is.

Or take the attitude of planning ideally to start writing that lecture, while at the same time you plan mundanely to read the newspaper. No doubt you are familiar with the mundane plan and can identify it easily. It is a disposition that will probably cause you to read the newspaper. On the way, it may cause you to get out of your chair, go to collect the newspaper from the table, and so on. But are you familiar with the ideal planning attitude that, in this case, is to start writing that lecture? Once again, I think you will only be able to identify it as your cognitive attitude of believing you ought to start writing that lecture. Likewise with your attitude of planning not to conspire against Caesar, if you are Brutus—I suspect you will recognize that attitude only by recognizing it as your belief that Brutus ought not to have conspired against Caesar.

This is how I think you are going to have to identify the ideal planning attitudes that Gibbard is talking about. You will have to recognize them as normative beliefs. This method will not steer you wrong. So far as Gibbard is concerned, you will identify the right attitudes this way, since ideal planning attitudes (provided you do indeed have them) are indeed normative beliefs. They have both the property of being planning attitudes and the property of being normative beliefs. So I am not contradicting what Gibbard is saying.

However, I do think this point about identification puts in doubt Gibbard's project of explaining our normative beliefs. The underlying noncognitive attitudes are supposed to explain how we have the cognitive attitudes. More accurately, an attitude's property of being a planning attitude is supposed to explain how it has the property of being a normative belief. But we can only recognize the underlying noncognitive attitude by recognizing its property of being a cognitive attitude. Gibbard in effect tells us that the belief that you ought to start writing that lecture is explained by the noncognitive attitude, whatever it is, that you recognize as the belief that you ought to start writing that lecture.

The explanans is identified through the explanandum, and this happens extensively throughout the explanatory story. This makes me doubt that we are being given much of an explanation at all. For a proper explanation, we should have some independent means of recognizing the explanans.

More particularly, my doubt is this. Gibbard's idea is that the noncognitive, planning attitudes, provided they are rational, are supposed to be woven together in a structure that explains how they can be treated as beliefs. But now it emerges that these attitudes can be identified in the first place only through their derived property of being beliefs. This means 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. Gibbard's story is that they stand in these relations because they are rational planning attitudes, but actually it is because they are rational beliefs. It is not that planning attitudes earn the right to count as cognitive attitudes; they have this right because we identify them as cognitive attitudes.

Comment on Lecture III

Gibbard gives great credit to Harsanyi in developing his utilitarian version of contractualism. He reminds us that Harsanyi in the 1950s proved two distinct theorems that can be used to support utilitarianism. The first makes use of the ideas that Rawls later called 'the original position' and 'the veil of ignorance.' So that

theorem is directly in the contractualist tradition, but the second is not. Nevertheless, Gibbard recruits the second theorem as well as the first to support his contractualist position.

One special feature of Gibbard's contractualism is that he thinks morality requires us to settle on a common goal, which each of us should pursue. He objects (p. 64) to a moral theory that allows each person to pursue her own distinct goals. He points out that, if we do each pursue our own goals, we shall encounter prisoners' dilemmas. The effect will be that everyone ends up satisfying her goals less well than they would have been satisfied had we all cooperated in pursuing common goals. Gibbard says (pp. 66–67):

> Whatever reasons each has for the peculiarities of her own goals, there is a way better to advance, in prospect, all these goals at once. The way is to agree on a common scale of goals for all to pursue.

Gibbard uses Harsanyi's second theorem to support this claim. He also uses it to support a second, subsidiary claim that this common goal is utilitarian in a broad sense: It is a weighted average of the goals of individuals (p. 67).

To be more accurate, Gibbard supports these claims, not with Harsanyi's own theorem, but with what he calls a 'Harsanyi-like' theorem. His appendix describes this theorem and illustrates it in a diagram. My figure B1 is a copy of this diagram. The axes show values of the variables v_i and v_j, which are two people's 'goal-scales'; they measure the degrees to which the people's goals are satisfied. Each point in the diagram marks a combination of values for v_i and v_j. The curve is the frontier of the set of points that are feasible; points on or below this frontier are feasible; points above it are not. We assume that one of the feasible points is ideal. We assume a 'Paretian' condition for this ideal. That is to say, we assume an arrangement is not ideal if it is possible to better satisfy one of the people's goals without satisfying the other person's goals less well.

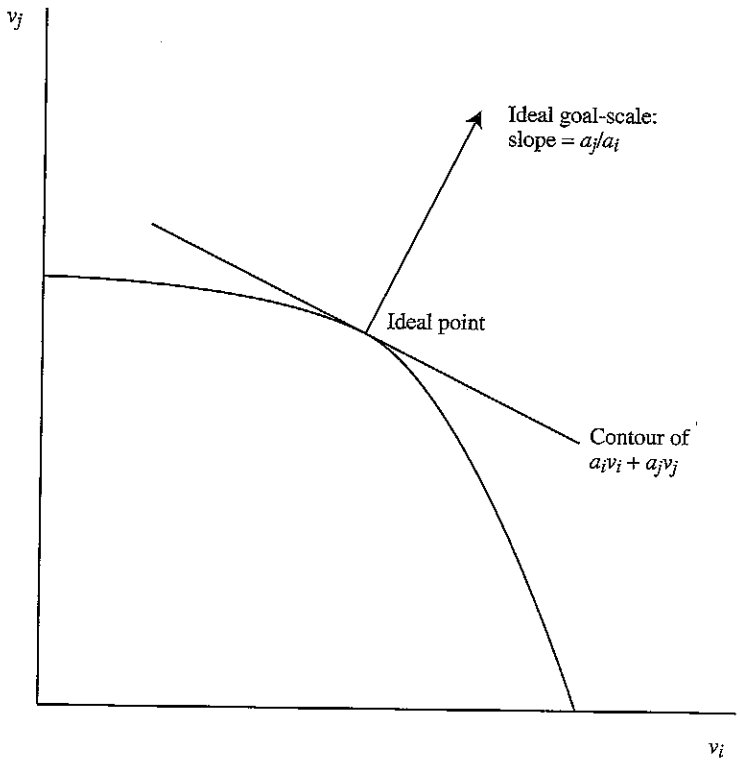

Figure B1

This ensures that the ideal point is on the frontier. We also assume that the set of feasible points is strictly convex, which means that the frontier bows outward.

Given these assumptions, the theorem Gibbard appeals to says that there is some weighted average $(a_iv_i + a_jv_j)$ of v_i and v_j such that the ideal point maximizes this average among the points in the feasible set and is the only point to do so. The tangent in figure B1 is a contour of this weighted average. A line perpendicular to the contour has slope a_j/a_i. Gibbard calls the weighted average the "ideal

goal-scale." The theorem generalizes from two to many people, but this two-person example is enough for what I need to say.

I do not think Gibbard should call this theorem 'Harsanyi-like'; I see little resemblance between it and Harsanyi's second theorem. It is one of the elementary theorems of convex analysis. I shall call it 'the tangent theorem.' As Gibbard points out, geometrically the theorem simply says that, at any point on the frontier of a strictly convex set, a tangent can be drawn that meets the set at that point only.

Moreover, I do not think the tangent theorem offers any worthwhile support to Gibbard's view that we should pursue common goals. It tells us that the ideal point could be achieved by maximizing the ideal goal-scale Gibbard defines. But it is a big step from there to the conclusion that we should pursue common goals. Gibbard himself mentions one difficulty in making that step. Even if we all independently pursue common goals, we may together fail to achieve those goals, because we may fail to coordinate our individual actions properly. But, as he explains (p. 85), Gibbard himself long ago proved a theorem that overcomes this difficulty in some circumstances. It is not this difficulty that concerns me.

The difficulty that concerns me is that, for all the tangent theorem tells us, the ideal goal-scale may depend on the shape of the feasible set, and on where in the feasible set the ideal point is. Alterations in the feasible set will change the ideal point, and we have no reason to think they will not alter the ideal goal-scale. We therefore cannot know what the ideal goal-scale is until we know what the feasible set is and which point in it is the ideal one. But in view of the complexity of life, we cannot possibly know the shape of the feasible set, and we certainly cannot know the position of the ideal point.

Figure B2 illustrates. It shows the frontiers of two possible feasible sets. I have picked out ideal points on each of these frontiers. Our theorem tells us nothing about where they are, so I have picked them arbitrarily.

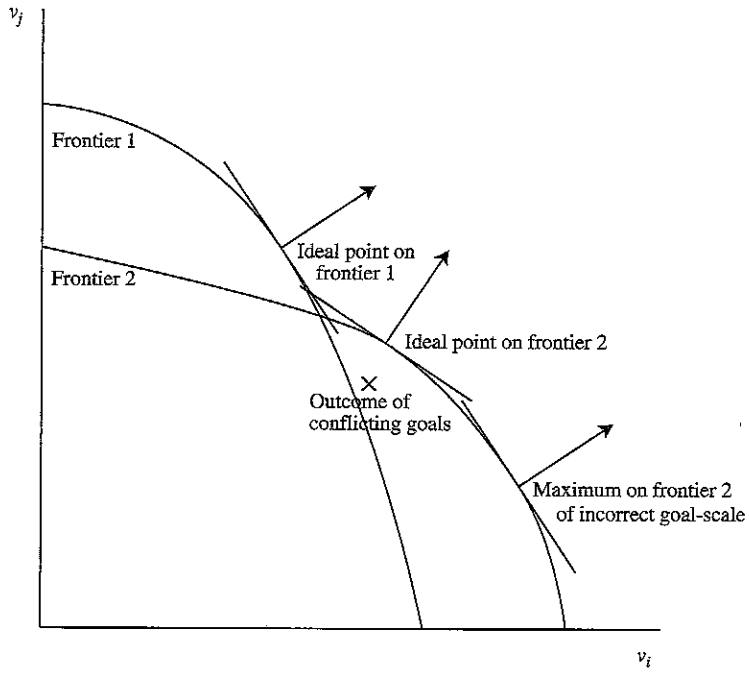

Figure B2

The two I have picked would be achieved by maximizing different ideal goal-scales within their respective feasible sets. We have no way of knowing which is the right goal-scale to maximize. We might easily get it wrong, averaging the two people's individual scales using the wrong weights. Then, if we maximized our wrong scale, we might end up far from the ideal point. Indeed, we might end up in a worse position than we would have reached if the two people had separately pursued their own goals. To be sure, we would end up somewhere on the frontier rather than inside it, as figure B2 shows. And to be sure too, if the people pursued their own disparate goals, they might end up somewhere inside the frontier, because of bother with prisoners' dilemmas. Nevertheless, the point they end up at might be better than the one achieved by maximizing

the wrong goal-scale. Inevitably, some points inside the frontier are better than some points that are on the frontier.

We therefore cannot conclude that we necessarily advance our goals better by choosing common goals to pursue.

For a different reason, the tangent theorem also does not support the subsidiary claim that, if we do choose a common goal-scale, it should be a weighted average of individual goal-scales. True, it tells us that the ideal point can be achieved by maximizing a weighted average of individual goal-scales. But there are many other functions such that the ideal point can be reached by maximizing one of them. For example, figure B3 shows it can be reached by maximizing a minimum function, specifically the function $\min\{(v_i - v^*_i), (v_j - v^*_j)\}$, where v^*_i and v^*_j are the values of the individuals' goal-scales achieved at the ideal point. Indeed, maximizing this sort of minimum function has an advantage over maximizing a weighted average. Figure 3 shows it will work even when the feasible set is not convex.

All in all, the tangent theorem is far too weak to give worthwhile support to any sort of utilitarianism. It will not do what Gibbard asks it to do. But Harsanyi's second theorem is very much more powerful; indeed, its conclusion is remarkable. Here is a statement of it. To match Gibbard's purposes, I have interpreted it in terms of goals and goal-scales.

Assume:

1. Each person's goals are coherent (which means they satisfy the axioms of expected utility theory).
2. The common goals are coherent.
3. The common goals satisfy the 'Paretian' condition: that if each person's goals are indifferent between two prospects, then the common goals are indifferent between those prospects; and if one person's goals place the first of two prospects above the second, and no person's goals place the second above the first, then the common goals place the first above the second.

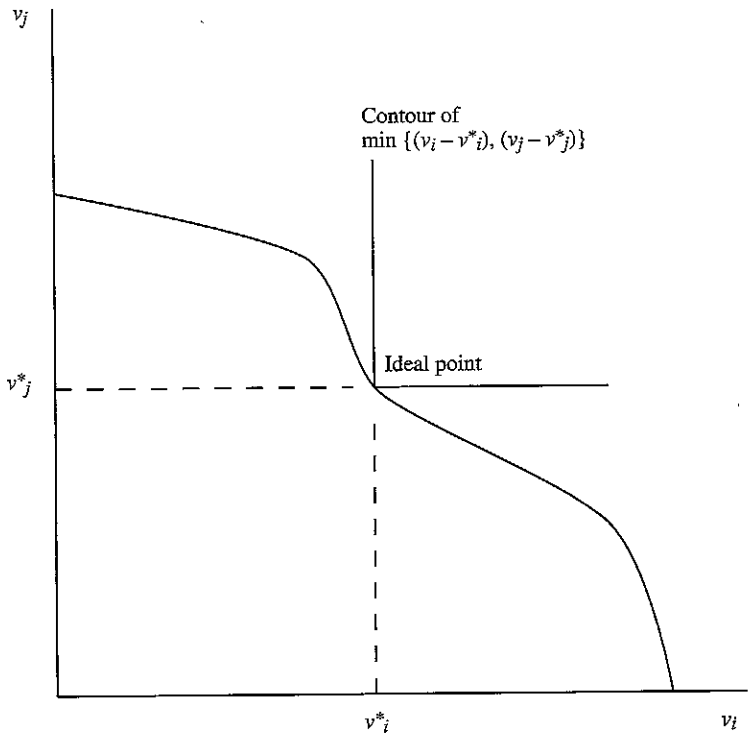

Figure B3

Then:

The common goals can be represented by an expectational goal-scale that is a weighted average of expectational goal-scales that represent the goals of the individuals.

(A goal-scale is said to *represent* a set of goals if and only if, whenever the goals rank one prospect at least as high as a second, the first has at least as high a value on the goal-scale as the second. A goal-scale is said to be *expectational* if and only if the value it assigns to a prospect is the expectation of the values it assigns to the prospect's possible outcomes.)

This theorem is illustrated in figure B4. It tells us that all points in the diagram can be ranked by the common goals in a way that is independent of the feasible set and of the ideal point. It therefore gives the people something they can agree on without knowing the feasible set or the ideal point. Whatever the feasible set turns out to be, the ideal point will always be whatever point in that set maximizes the common goal-scale. Moreover, the theorem tells us that these common goals are a weighted average of the individuals' goals.

This is just what Gibbard needs. So in the end, I think he is right to say that Harsanyi's second theorem gives him what he needs, but wrong to suggest that a weaker surrogate will do.

Moreover, this theorem would give him more than he seems to realize. He says (p. 70), "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...." Well, Harsanyi's theorem establishes it. That is why this theorem is remarkable. One of its premises is that each person's goals are a consideration; each person's goals count. That is what the Paretian condition says, in effect. Simply on the basis of the coherence conditions and this assumption that each person's goals count, the theorem concludes that they count specifically in an additive fashion. Their weights are added. The theorem derives additivity from those remarkably weak premises. Gibbard does not need to assume additivity; he could take it from the theorem.

Furthermore, the theorem answers a question Gibbard raises at the end of the appendix. He says (p. 87), "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... all maximize the same goal-scale." The answer from Harsanyi's theorem is: "Yes, they do." As I explained, the goal-scale is independent of the feasible set of prospects.

Given all the merits of Harsanyi's own theorem, why did Gibbard eschew it and instead fall back on a theorem that turns out too

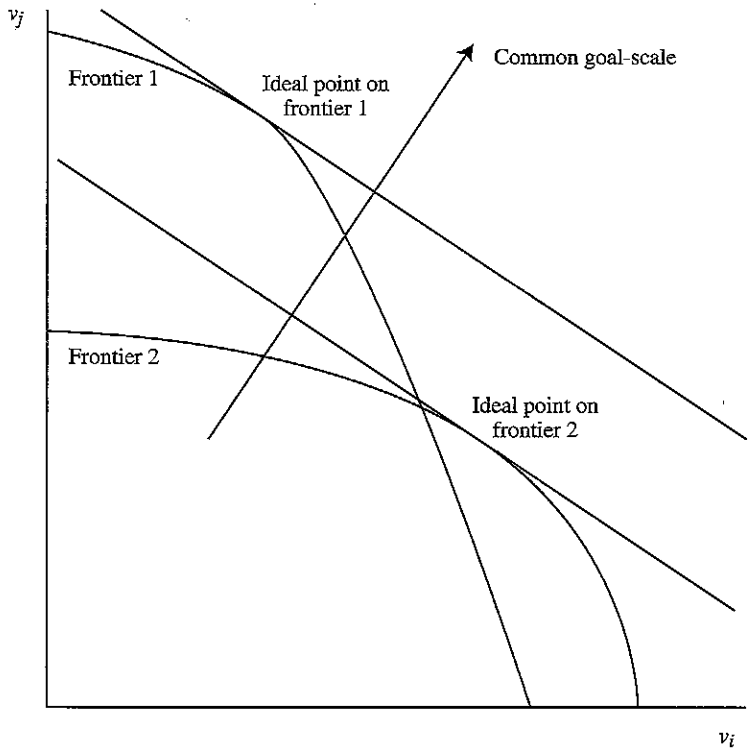

Figure B4

weak for his purposes? His answer is explicit on p. 65. It is that he thinks the second premise of Harsanyi's theorem—that the common goals are coherent—is open to question. But I have explained that he needs this premise. He cannot happily give it up and fall back on a weaker theorem, because the weaker theorem is not up to the work he demands from it. I therefore think he needs to try to establish the coherence of the common goals. If the common goals are indeed not coherent, that will do serious damage to the argument of his Lecture III. It will weaken his case for a utilitarian type of contractualism.

I agree with Gibbard that the coherence of the common goals is open to question. Nevertheless, the answer to this question might be that they are. I do not know. I do have my own arguments in defence of the second premise of Harsanyi's theorem; they are set out in my book *Weighing Goods* (Blackwell, 1991). But I interpret the theorem differently from Gibbard—in terms of good rather than goals. I take it to be telling us something about the structure of good, and specifically about how the overall good is related to the good of individuals. Under this interpretation, I believe Harsanyi's theorem can be used to give strong support to utilitarianism, or more exactly to a utilitarian theory of value. Under my interpretation, the theorem's first premise is that the good of each individual is coherent, and the second premise is that overall good is coherent. I believe these premises can be convincingly defended. In particular, I believe that overall good is coherent.

But Gibbard does not wish to interpret the theorem in terms of good. For one thing, doing so would not be so conducive to his contractualism. Furthermore, in view of T. M. Scanlon's attack on the notion of individual good, he does not wish to rely on this notion at this point in his argument (p. 66). So he turns instead to the more general notion of an individual's goals. Correspondingly, he turns to common goals instead of overall good.

However, he does actually accept a notion of overall good, and he takes it for granted that the common goals achieve overall good. "By the overall good," he says (p. 67), "I shall mean good as measured by whatever goal-scale would be specified by the social contract." If we may take this assumption seriously, then my own arguments for coherence will apply to the common goals, because they apply to overall good. So I see some prospect of justifying the second premise of Harsanyi's theorem, even under Gibbard's interpretation. I cannot say more than that, because much more work would be needed to develop a full defence. It would have to be explained how we can be sure that the common goals do indeed achieve something that can count as overall good.

My point is that Gibbard cannot satisfactorily evade this work. Though he does not realize it, for his purposes, he needs the common goal-scale to be coherent. He needs to use the genuine version of Harsanyi's second theorem; a weaker substitute is not enough.