Allan Gibbard

Matthew D. Adler and John A. Weymark

This interview was conducted on October 15, 2015 at the home of Allan Gibbard (**AG**) in Ann Arbor by Matthew Adler (**MA**) and John Weymark (**JW**). Adler participated by Skype from his office at the Duke Law School. The text of the interview has been edited to improve its readability, to clarify some of what was originally said, and to provide bibliographic details for the works cited.

JW: Allan, let us begin with some questions about your background. You grew up in West Virginia where your father, Harold Gibbard, was a prominent Professor of Sociology. Did growing up in an academic family play an important role in your own decision to become an academic?

AG: Well, the psychologists tell us that we don't know the causes of our actions just by introspection, but I'm sure it had an important influence. I grew up with academic values, and for a long time I thought that being a perpetual student was the ideal life.

JW: You went to Swarthmore College as an undergraduate where you majored in Mathematics and minored in Physics and Philosophy. Why did you choose this combination?

AG: Well, I had a wonderful introductory Philosophy teacher, Jerome Shaffer, and had no idea I was interested in Philosophy before that. I thought that philosophers specialized in giving fallacious proofs of the existence of God. I'd gotten fascinated

M. D. Adler
Duke Law School, 210 Science Drive, Durham, NC 27708, USA
e-mail: adler@law.duke.edu

J. A. Weymark (✉)
Departments of Economics and Philosophy, Vanderbilt University, VU Station B #35189, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA
e-mail: john.weymark@vanderbilt.edu

M. Fleurbaey and M. Salles (eds.), *Conversations on Social Choice and Welfare Theory - Vol. 1*, Studies in Choice and Welfare,
https://doi.org/10.1007/978-3-030-62769-0_11

with Physics, I think in 5th grade, but then laboratory Physics didn't work very well for me, so Mathematics was more down my line, although I don't think I would have been a good enough mathematician to thrive as a professional mathematician.

JW: In your 2006 interview with Alex Voorhoeve (Voorhoeve 2009), you mentioned that both Shaffer and Richard Brandt were influential. Later, we'll talk about how Brandt influenced your choice of Ph.D. thesis. Are there other ways that Philosophy faculty at Swarthmore influenced the kind of issues you considered?

AG: Well, Brandt was the most important figure for me, and he had a very direct way of trying to figure out what's really at stake with a problem that people were discussing, and that very much appealed to me. I didn't study with him until the Fall of my senior year when I took his Moral Philosophy seminar, although my freshman year roommate was in his introductory class and we discussed things a great deal. So, the appeal of Brandt's approach to Philosophy was just very strong for me.

MA: This is just a quick follow up. Brandt was later at Michigan and you, of course, have been at Michigan for a while. Were those connected?

AG: Yes, I think so. Brandt left Swarthmore the year after I graduated, and during the era when he was there, Swarthmore had one of the top departments and it was really built by Brandt. And then in the 1960s universities came to dominate colleges, so there weren't going to be college departments of the eminence of the Swarthmore one by the time the 1960s came. But, it was remarkable when I came to Michigan that two other members of the department had been teaching as young Assistant Professors at Swarthmore when I was there and one other member of the department had been a student at Swarthmore when I was there.

MA: Who were those people?

AG: Larry Sklar and Jaegwon Kim had both been Assistant Professors at Swarthmore. I got to know Larry there, but never took a class from him. But I took a Symbolic Logic course from Kim. John Bennett was the one who I think was a freshman at Swarthmore when I first got to know him.

JW: You have partly answered the next question because in your later work in social choice theory you employ formal logic and set theory. You have already explained how you learned about logic as a student. Did you also learn about set theory from your days at Swarthmore?

AG: Yes, I think the Swarthmore program was much better then than it was later when my son went there, because there was a real interest in the abstract aspects of Mathematics. Those of us who had finished second year calculus as freshmen were put in a course that studied things like the Peano axioms and set theory and various other sorts of highly abstract aspects of Mathematics and that was the part of Mathematics that fascinated me.

JW: Following your graduation from Swarthmore, you spent two years with the US Peace Corps as a high school teacher in Ghana. What prompted you to do this and has this experience had an influence on your research?

AG: Well, I was, like everyone, worried about the balance of my obligations to others and doing the things I wanted to do. I didn't think I could be an Albert Schweitzer who did the things he wanted to do until he was thirty and then became a doctor in Gabon, but I decided I could spend two years trying to be of service before I went on to do the things I most wanted to do. I discovered in the Peace Corps that virtually nobody else in the Peace Corps had such a motivation, or would at least avow it. They were of course quite devoted to their teaching. For me as well there were great benefits to being in the Peace Corps. I'd never been outside North America, and getting to know another culture broadened my view of human possibilities in an important way. So, I got a lot of personal benefit from being in it.

JW: Do you think it influenced how you approach moral issues?

AG: Yes, I think so. I wrote a paper for John Rawls in graduate school where I was trying to argue that there are lots of people who don't have the kinds of goals that Rawls attributes to his parties in the Original Position (Rawls 1971). I was very struck that, whereas I had grown up in an academic family where the highest value was having a free mind and enquiring into things, in Ghana the highest values were behaving well and being obedient. Of course, you can get that within American culture too. I had grown up in a small slice of American culture. Indeed it occurred to me later that I'd grown up in a colonial situation, and that the university was full of people who weren't from West Virginia or similar places but were surrounded by people who were. I imagine that's the way it is where you are too.

MA: Allan, was the concern there that people would have different conceptions of the personal good or was it rather more of a concern like Tim Scanlon's that people behind the veil would be motivated by altruistic considerations as well as considerations of personal good (Scanlon 1998)?

AG: I always thought that people would be motivated by altruistic considerations as well as by their own good, but Rawls takes freedom to form and revise one's plans as a top value that the social contract is to promote. The people I met in Ghana may have had this view of their own lives—I'm not sure—but it didn't seem to be part of their child-raising, as it is among the American liberals I know. Good behavior and obedience seemed to be what was stressed.

MA: Right. But does one get around that by looking into fully informed—ideal—preferences? Of course, ideal preferences are something you talk about a lot in *Wise Choices, Apt Feelings* (Gibbard 1990b) and *Thinking How to Live* (Gibbard 2003).

AG: Well, the people in Ghana certainly had preferences, but as I say, it didn't seem that they had been formed in an ideal way by free-thinking about what to want. That didn't seem to loom very large for people who were concerned that their children were going to grow up well behaved but not concerned that their children grow up with enquiring minds. This isn't to say that people didn't have enquiring minds, but it didn't seem to be the same sort of value as it was among the academics I'd grown up with.

JW: Let us move on to the next stage of your career. In 1969, you started graduate school at Harvard in Philosophy. In your first term, you took a course with John Rawls in which you read parts of the manuscript of his *A Theory of Justice* (Rawls 1971). How did Rawls' book influence your own views about these issues then and subsequently?

AG: Well, his argument that altruism is not the strongest motivation, that some sort of reciprocity is a stronger motivation, made a big impression on me. And it was a very exciting class. Rawls was thinking of the Original Position as a site for bargaining, and one day some of us stayed after class and argued with him for about an hour that people in the Original Position wouldn't have clashing interests that they knew about, so that the bargaining wasn't really genuine bargaining. I think that's the position that Rawls took eventually. All that was very exciting. The importance of the idea of fair reciprocity was one thing that struck me from being exposed to Rawls that I hadn't been struck by being exposed to Brandt.

JW: Willard Quine and Hilary Putnam were also on the Harvard faculty at that time. Did your interactions with them or any of the other Harvard Philosophy faculty play a major role in shaping your views about philosophical issues?

AG: Well Quine had a big influence. I didn't end up having all of his views, but he was sort of the dominant figure, and at Swarthmore I was the sort of logical positivist he criticized. Swarthmore was very caught up in the ethos of the, say, late 1930s to 1950s in English analytic philosophy. I remember reading Quine's "Two dogmas of empiricism" (Quine 1951) sitting under a tree and feeling quite dismayed and then trying to figure out how to integrate those sorts of views into my view of the subject. My most recent book, *Meaning and Normativity* (Gibbard 2012), is really trying to see how much of the older approaches one could save if one thought that questions of meaning were normative questions and not purely empirical questions.

And Putnam, well, Putnam was a sort of avid communist. Engels was I think his favorite philosopher, and I had had enough exposure to what communism was like to be resistant to that. So I listened to Putnam a lot, but I certainly didn't become a follower. I guess it was later that he had his transformation that resulted in his transformative philosophical article, "The meaning of 'meaning'" (Putnam 1975), and at that time he gave up his hard-line communism. I'm not sure whether it was Maoism or what; most of the graduate students were enthusiastic Maoists. Putnam also taught a course on the advanced logic of the Continuum Hypothesis and things like that. I learned a lot from that.

JW: I understand that your 1971 Ph.D. thesis, *Utilitarianism and Coordination* (Gibbard 1990a), is a subject that you started thinking about when you were still at Swarthmore. Can you tell us a little bit about the development of the origins of your thesis and what the basic ideas in the thesis were?

AG: I took Brandt's Moral Philosophy seminar in my senior year, and that was, I suppose, the biggest undergraduate influence on me. And one week the subject was rule utilitarianism and variations on it. Brandt each week would have a list of articles

that were central to the subject for the week, and each student would every two weeks write a short paper. This was my week to produce a paper, and as he was talking about the readings for that week, he said, "Oh, and I have a little thing that I've placed on reserve" (Brandt 1963). But somehow I forgot about Brandt's "little thing" and read the other papers and formed a view and wrote my paper. And then Brandt said, "Oh, did you get a look at my 'little thing'?" And I turned red and realized I'd forgotten it, and so I felt I had to write another paper. But by then I'd formed independent views on the issue his paper concerned. He argued in his paper that a fairly straightforward form of rule utilitarianism would be what he called "extensionally equivalent" to act utilitarianism. I had concluded in my work before reading his article that they weren't, that there were situations where the question "What if everyone did the same?" made a real difference. So, I wrote the second paper and Brandt said, "Well, if this is true, it ought to be publishable." Of course that's an exciting thing to hear as an undergraduate. So, I worked on a revised version, and he read it and said, "If this is true, it ought to be publishable. Of course, it isn't publishable yet." And then I had to do a few iterations.

When I was in Ghana I got a letter from him saying that a young genius named John Troyer in his seminar had produced a paper on my paper, and Brandt wrote a two-page letter along with Troyer's paper. I thought Brandt and Troyer were both getting matters completely messed up, and I decided that the only way to write this paper was to pretend I was talking with my very intelligent little sister when she was twelve. I was fifteen when she was twelve, and in that period, I was explaining lots of things to her, and so I had lots of practice at explanation with immediate feedback on whether I was getting things across. And so imagining explaining things to her produced a paper that everybody liked, and it was accepted by the *Australasian Journal of Philosophy* (Gibbard 1965) before I got to graduate school.

I had a four-year fellowship at Harvard, and most people took longer than four years to finish the Ph.D. But if I took longer than four years, I thought, I was going to have to teach "Hum 5," the course Humanities 5, which is a very broad survey of so-called western philosophical thinking, and it seemed to me that teaching that course was going to be a full-time job. I wouldn't be making any progress, so I was very interested in finishing up in my fourth year. I spent a summer trying to figure out a thesis on ethical relativism and came up without much to say. And then I got a further idea of how to develop the arguments about the non-equivalence of rule utilitarianism and act utilitarianism.

And there was also a book by a man named D. H. Hodgson (Hodgson 1967), who was an Australian judge, and he was arguing that ideally rational act utilitarians would not be able to make and keep agreements, because the motivation to keep the agreement depends on the expectations that others will be keeping it. I got interested in that. I don't think I'd ever heard of a Nash equilibrium at that point. David Lewis, though, was writing his book *Convention* (Lewis 1969), drawing on game theory to produce a theory of meaning. Lewis had worked for the Rand Corporation, I think, and he knew a lot about these things. So I developed another part of the thesis that addressed Hodgson, exploring what kinds of agreements people would keep if they could establish the expectation that the agreement would be kept. So that made a

dissertation. Those weren't the subject matters I expected to be pursuing later, but the thesis did get me through without having to teach Hum 5.

MA: I'm struck by the fact that so many people, great minds, you and others, worked on this problem of the divergence or convergence of act and rule utilitarianism and the related question of coordination among utilitarians. As you mentioned Lewis did so, and of course David Lyons (Lyons 1965), Hodgson—and I believe your colleague Don Regan, who now teaches at Michigan Law School, also wrote a book on a similar topic (Regan 1980). And then of course, Derek Parfit later in the first part of *Reasons and Persons* (Parfit 1984) talks about this and I'm struck by the fact that there's not been subsequent scholarship; it sort of dies out. I wonder, what are your thoughts about that? Is it that so much attention then shifted to Rawls; rather than looking at the different variants of utilitarianism, the focus after Rawls is on the debate between justice and utilitarianism or consequentialism? Anyway, it is striking to me that there was so much work on this (the divergence or convergence of act and rule utilitarianism), including yours, between, 1950 and the mid-1980s and since then there seems to be less scholarship.

AG: Yes, that's a good question and I don't entirely know what to say. But it seems to be that what you suggest is right: There is a sort of vast culture shift from thinking that utilitarianism basically had it right and one had to work out some problems, to thinking that the right view was somehow contained in Kant and that utilitarianism had it all wrong. I guess I was never convinced that the truth was all contained in Kant. Rawls had a tremendous influence on this shift, I think.

JW: Speaking of Rawls, we would like to hear something about your experience having Rawls as your thesis supervisor. While you both address issues related to how a community should coordinate activities so as to realize the benefits of cooperation on reasonable terms, you were considering broadly utilitarian principles, whereas Rawls rejects utilitarianism in favor of a form of liberal egalitarianism. Does this difference in views affect how you interacted? Do you think that having to defend your ideas to Rawls helped you to refine and sharpen your arguments?

AG: Well, we interacted on working out his views and implications of it. He wasn't very interested in the subject of my thesis and so the advice amounted to: "Yes, I read it. It's very good. Some parts of it are a little obscure." "Oh, what's obscure? I'll try working on that." And then he gave an example, so I rewrote that part. He claimed I didn't need much supervision and certainly I didn't get much supervision.

JW: Also, during that period, you have recounted that you first came across Ken Arrow's *Social Choice and Individual Values* (Arrow 1951) in the Harvard Philosophy department's Robbins Library.

AG: Yes. It's a good thing that it has a bright blue cover.

JW: Do you recall what your reaction to reading the book was?

AG: Oh, I was just amazed and very puzzled. And then when I heard that Arrow was actually coming to Harvard and that he was collaborating with Rawls on a seminar

with a young economist I'd never heard of named Amartya Sen, I was very excited, and I told Rawls I wanted to take it. He said, "Oh that's just for Economics graduate students." But I insisted—and the seminar was of course an utterly amazing experience. Howard Raiffa was there, Franklin Fisher, his co-author Jerome Rothenberg, and a young man people said was national bridge champion—Richard Zeckhauser. I sort of sat there gaping and thinking, "I've never witnessed such intelligence all gathered in one room before."

JW: In that course, Sen circulated drafts of some of the chapters of *Collective Choice and Social Welfare* (Sen 1970a). Other than Arrow's book, had you read anything else in social choice theory before this? How did Sen's book influence you?

AG: Sen's book was sort of encyclopedic. I don't think it transformed my view of the subject, but it certainly had things worked out marvelously. But I was still focused on what are we to make of Arrow's Theorem. That's what I—well, you've read the seminar paper that I ended up writing.

JW: O.K. Well, let me ask about that. The term paper you wrote for that course won the Goldsmith prize that year for the best paper in an Economics course or seminar and it's recently been published in *Economics and Philosophy* (Gibbard 2014b). In it, you established your well-known oligarchy theorem: If Arrow's collective rationality condition is weakened to quasi-transitivity while maintaining the rest of his axioms, then the social choice procedure is oligarchic, what you call a "liberum veto oligarchy." What motivated you to consider weakening Arrow's transitivity assumption and why did you use the term "liberum veto oligarchy?"

AG: Well, Sen had devoted a session to advocating weakening transitivity to what he called quasi-transitivity. So, I don't remember how I discovered that that wasn't going to be much help, but in the seminar paper I was addressing Sen. I'd read a lot of history growing up, including Polish history, so the liberum veto was part of my background knowledge and seemed like the right term to adopt for it.

JW: When you subsequently rewrote this part of your term paper around 1970 while an Assistant Professor of Philosophy at the University of Chicago, you recast it more formally as a response to a paper by Schick (1969) that argued that Arrow's Impossibility Theorem is not particularly troubling because the transitivity of social preferences is, in his words, "untenable." Why did you rewrite your theorem in this way and why did you not publish the paper then? It was subsequently published quite recently in the *Review of Economic Design* (Gibbard 2014a).

AG: Well, the problem I had with the seminar paper was that I was worried there was nothing to respond to that was in print. Sen's lecture on the subject had been, as far as I knew, a lecture to the seminar. So I didn't quite know how to place the significance of the theorem. I guess I shouldn't have been so diffident, but when Fred Schick's paper came out, then there was a clear target. So my theorem spoke to something that someone had actually advocated in print.

MA: Let me just follow up quickly on transitivity. As you probably know, there's been more recently a debate in moral philosophy about transitivity. Larry Temkin has

argued for giving up transitivity in his big new book *Rethinking the Good* (Temkin 2012). He argues for giving up not just transitivity of indifference, but transitivity of strict preference. On the other hand, people like John Broome (Broome 2004) have said that transitivity is just analytic in "betterness." Do you have a view about that? Whether transitivity is just sort of built into consequentialist thinking or not?

AG: I think my view would be much more Broome's. I haven't worked afresh on it, but it's always seemed to me that the transitivity of "better than" is obvious.

JW: My understanding is that you actually never submitted that paper. Can you tell us why?

AG: Oh, well, Hugo Sonnenschein sent a paper, his joint paper with Andreu Mas-Colell (Mas-Colell and Sonnenschein 1972). And then there was another paper by Guha (1972). So the result had been proved and there didn't seem to be a lot of point in duplicating those proofs.

JW: In your Harvard term paper, you also consider the implications of dropping Arrow's independence axiom. Using a version of your well-known Edwin, Angelina, and the judge example that subsequently appeared in your 1974 *Journal of Economic Theory* article (Gibbard 1974), you showed that the Pareto condition is incompatible with a particular kind of liberal right—the right of two people to get married should they want to. As in Sen's Impossibility of a Paretian Liberal Theorem (Sen 1970b), a right links individual and social preferences by requiring the social preferences over a pair of alternatives to coincide with an individual's preference when it lies within his private sphere. Although Sen had presented his Paretian liberalism theorem to the Econometric Society before going to Harvard, am I correct in understanding that he did not discuss this result or circulate that part of his book in the seminar and that you were unaware of it when you wrote your term paper? Also, what led you to consider modeling rights in a social choice framework?

AG: Well, if I can remember this, I wasn't aware of Sen's paper, although maybe I should have been. But as I seem to remember, after I presented my paper, he gave me a paper. I was somewhat confused about its status and didn't get that straight until much, much later. So I think when I presented the theorem to the seminar, I didn't know that Sen had done that.

JW: What motivated you to try and model rights using the social choice framework that Arrow had developed?

AG: I don't remember in specific terms. It did seem that one could use the social choice framework to talk about rights, and that then important and interesting things happened, including things that bore on what the significance of the Arrow Theorem was and wasn't. I don't remember having an epiphany that rights could be treated this way, but it certainly seemed that they could be.

JW: In that term paper, you suggested that the way you modeled rights is not completely satisfactory and raised the possibility, but didn't explore, that it may be desirable to waive a right. These are issues that you explored in your 1974 *Journal*

of Economic Theory article (Gibbard 1974). In that article, you modeled a set of alternatives as the Cartesian product of the personal features that are available for each individual. Using this framework, you showed that a natural way of assigning rights is inconsistent, without any appeal to a Pareto principle. You also allowed individuals to waive rights and showed that this formulation of rights is both self-consistent and compatible with the Pareto principle. What advantages does this way that you modeled alternatives and rights-assignments have over the way that they were modeled by Sen?

AG: Well, Sen was trying to show that the Pareto principle was defective, and his argument was that the Pareto principle was inconsistent with assignments of rights. That seemed very puzzling to me, since even though, as Sen formulated things, the awarding of rights was not inconsistent without bringing in the Pareto principle, it seemed that the motivation for thinking that people had rights was a principle that was going to turn out to be inconsistent by itself. So I didn't understand how what Sen was doing was going to discredit the Pareto principle. I think Sen has always stuck to the view that it does, and I'm as puzzled as I ever was.

JW: In *Anarchy, State, and Utopia* (Nozick 1974), Robert Nozick took issue with the way that both you and Sen model rights. In your approach, rights are linked to preferences—when an individual has a preference over a pair of alternatives in his private sphere, then that is the social ranking unless the right is waived. Nozick instead argued that by exercising their rights, individuals pick some features of an alternative and that social choice considerations only apply when determining any remaining features. How would you respond to Nozick? Do you still think that your way of formalizing rights in a social choice framework is a good way?

AG: Social choice theory concerns what histories are morally O.K. in the ways that what happens depends on people's preferences. And that seems like a question one can pose when one thinks about rights. And if you pose it that way, then the social choice theory apparatus applies.

JW: How do you respond to Nozick's way of thinking about rights, which is more game theoretic? In his approach, each person independently chooses within his or her own sphere and then there may be some features left over, and that's the only part of the decision that's left to social choice considerations.

AG: Well, I don't see why one shouldn't apply the social choice framework anywhere where your assumptions are good assumptions, and so I don't think of the two ways of doing things as incompatible.

JW: Among economists, you are best known for what is now called the Gibbard–Satterthwaite Theorem (Gibbard 1973; Satterthwaite 1975). This theorem shows that any social choice function that maps profiles of individual preference orderings into a single choice from the set of available alternatives must be dictatorial if (1) the domain is unrestricted, (2) there are at least three alternatives, and (3) the social choice function is strategy-proof—that is, nobody can ever gain from misreporting his preferences. What prompted you to work on this problem? When you started

working on it, did you expect that the conclusions would be so nihilistic? And what challenges did you have to overcome to establish your theorem?

AG: As I remember it, Arrow, when we were discussing—"we" is the wrong term, because I don't think I said more than two things the whole seminar until I presented my paper—but when the seminar was discussing Independence of Irrelevant Alternatives, which I thought was what was problematic in the assumptions of the Arrow Theorem, Arrow said something about it being equivalent to strategy-proofness. And so the next year, I had started teaching at the University of Chicago and they had assigned me a social choice theory seminar which had, I think, two registered students and two auditors, including Bernie Grofman. Bernie was quite an auditor, and I was probably learning more from him than he was from me. Preparing for one session, I thought, "Well, I should work in the part about being immune to strategic manipulation being equivalent to IIA." I assumed it would take about five minutes to think it through. "So let's see, I'd better prepare this class. How does that work?" And then I got stuck and I was stuck for several weeks. And then I can remember the afternoon when I was visiting my fiancée Mary in Urbana and we decided to spend a couple of hours sitting at her card table getting some work done. The basic idea of how to prove what I was looking for occurred to me, and my mind started racing.

MA: Let me just add a quick follow up. You mentioned in passing there that you found Arrow's Independence of Irrelevant Alternatives to be problematic. Can you expand a bit on that? Why do you think that the axiom is problematic? This has been a big issue in social choice theory.

AG: Well, first, of course there's confusion in Arrow about what's governed by the Independence of Irrelevant Alternatives as he formulates it, and what is a matter of social choice being guided by a preference function that is fixed independently of what's feasible and what isn't. Once we get all that straight, the reason to reject IIA as Arrow formulates it is basically the one I gave in the seminar paper, that one's ordering of non-feasible alternatives, along with feasible alternatives, gives some indication of the strength of one's preferences. My example, I think, was Patrick Henry, "Give me liberty or give me death," which seems to indicate a strong preference for liberty. Arrow was a thorough ordinalist. In another version of the theorem, you have cardinal utility scales that aren't interpersonally comparable, and only ratios of preference strength matter. But either way, people's preferences involving infeasible alternatives provide some indication of preference strength. And so I thought the kind of information you can glean from preferences involving infeasible alternatives was relevant to the moral weight that a preference should bear.

JW: *Econometrica* was not where you first submitted your article. Please tell us something about what happened when you first sent your paper off for publication.

AG: Well, there was a new journal called *Theory and Decision* and I submitted it to that. And I was still puzzled that such a simple theorem didn't have a simple proof, and I included a covering note to the effect, "Maybe this is trivial, but I certainly can't see that it's trivial." I got a response from one Editor that said, if I remember,

"It is trivial. Arrow's Theorem says that any social welfare function is imposed or dictatorial. And so it is imposed or dictatorial or manipulable, and it is imposed, dictatorial, or a hot dog." So it was a completely fallacious, a complete misstatement of Arrow's Theorem. And then I got a two-page letter from the other Co-Editor who said the same thing in two long pages. So I decided I'd better try to send it to a competent place, and I sent it to *Econometrica*. I think I got indications later that Arrow and Sen had been the referees, and so there it was certainly competently refereed, to say the least.

JW: In order to establish your theorem, you introduced the concept of a game form, which specifies an outcome as a function of the strategies chosen by the individuals. In terms of a game form, strategy-proofness is then the requirement that truth-telling is a dominant strategy in the direct mechanism in which a strategy is a preference. Did you realize when you wrote your article that game forms would be a useful tool for analyzing other problems involving asymmetric information?

AG: Well, it seemed to me they were a useful tool. The first version of the theorem I proved for myself was the one that just talks about truth-telling and expressing one's preferences. It was later on, maybe in the summer of 1970 which I spent hanging around Cambridge University, that the game-form version occurred to me. I loved its generality. In lots of systems of voting, you aren't voting by doing something that constitutes making a statement of what your preference ordering over the alternatives is. At some point, it occurred to me that there's a more general way of putting the result. The game-form formulation has the advantage that it doesn't actually talk about truth-telling. It just says that for each preference ordering you might have, there's a dominant strategy, whether or not it takes the form of reporting your preferences truthfully. So that seemed obviously a much more general way to formulate things that would apply to any social mechanism whatsoever, and I was pleased with that—although of course it makes the statements of proofs more complicated.

JW: Mark Satterthwaite independently established a version of this theorem. His constructive proof first appeared in his Ph.D. thesis. In the published version (Satterthwaite 1975), in addition to providing this constructive proof, he also shows that the Gibbard–Satterthwaite Theorem follows from Arrow's Theorem and vice versa. In his article, Satterthwaite said that he got the insight for this correspondence result from reading your article. When did you first learn of Satterthwaite's work and did you have any inkling that your theorem and that of Arrow were so tightly linked before you read his article?

AG: My own proof proceeded via the Arrow Theorem, and that made for a tight link. A beautiful thing about Mark's proofs is that he got the strategy-proofness theorem independently of the Arrow Theorem, and then used all this to draw connections that I had not myself drawn. I remember getting Mark's letter when I was at Chicago—I think it was a letter, or perhaps he phoned. As I say, I didn't have a direct proof of the manipulability theorem—my proof went by way of Arrow's Theorem—and he did have a direct proof. I think he said that he had submitted it and

Sen had pointed out that I had proved the result, but then Mark did quite ingenious things with the implications of having a direct proof of the theorem.

JW: I guess a better way to phrase the second part of my question was that he seemed to sharpen the connection between Arrow's Theorem and the Gibbard–Satterthwaite Theorem, in effect showing that you can go back and forth between the two.

AG: Yes, right.

JW: In your article, you emphasize that the social choice is required to be deterministic. You later allow for the social choice to be a probability distribution over the alternatives (Gibbard 1977). You show that strategy-proofness implies that the social choice function must be a probability mixture of functions that are either dictatorial or only choose from two fixed alternatives. Do you regard such social choice procedures as being satisfactory? More generally, do you think that lotteries have a useful role to play in social decision-making?

AG: Well, I think obviously the narrow sorts of schemes that turn out to be nonmanipulable are not satisfactory, and one should tolerate some manipulability rather than have such blatantly unsatisfactory ways of choosing things.

JW: Lotteries are sometimes used when there are indivisibilities for which there is no way to split things up in some fair way between people. Do you think that that's a reasonable way to make decisions in such circumstances?

AG: Yes, I don't think that all possible uses of lotteries in social choice are unacceptable. Sometimes, a lottery is a reasonable way of resolving ties and the like. But I don't think that the indeterministic version of the theorem says, "Ah, there's a satisfactory way, there's a strategy-proof satisfactory way of making social choices after all." The kinds of indeterministic schemes that the theorem show can be strategy-proof are blatantly defective, in my view. And so you're still faced with trade-offs among desirable features, and in light of these trade-offs, full strategy-proofness will need to go.

JW: From what you said earlier about Independence, do you also have concerns that strategy-proofness is too strong of a requirement?

AG: We learn it can be had only at too high a price, so it's clearly too strong a requirement in that sense. It seemed like a nice feature in isolation, but the price for it is prohibitive. It seems to me that that's what the theorems show. After the impossibility theorems, people worked on what to do in the face of the finding that strategy-proofness can come only at an unacceptable cost. I didn't master the results that emerged, but that seems to me to be the right approach.

JW: Rawls' Difference Principle requires that social institutions be designed so as to maximize the prospects of the least advantaged as measured by an index of primary goods (Rawls 1971). This index is supposed to be a measure of an individual's command over basic social goods like income, rights, and opportunities. This

raises the problem of how such an index can be constructed. In your 1979 *Theory and Decision* article (Gibbard 1979), you tackled this problem using the formalism of social choice theory and welfare economics. You reformulated Rawls' Difference Principle in terms of a possibly partial ordering of the opportunities facing an individual. Among other results, you showed that if individuals have preferences over the relevant goods that satisfy the standard assumptions used in microeconomics, then your version of the Difference Principle is incompatible with the Pareto principle. What lessons for Rawls' theory do you think follow from this result?

AG: I haven't kept that paper well in mind over the decades, but I keep hoping that it offers the definitive interpretation of Rawls' Difference Principle, pursuing his motivations and avoiding definite mistakes. Rawls' system has a number of layers, but the crucial ones, to my mind, are these: First, the Original Position with the information and motives of parties who are choosing principles of justice to be accepted and realized in their society, and second, those principles themselves—which include the Difference Principle. I myself think that this system is at least consistent with a kind of indirect utilitarianism. Rawls, as far as I know, never precisely denied this; what he denied is that his system amounts to direct utilitarianism. I think that John Harsanyi (Harsanyi 1955) was right that the standard of what would be chosen in the Original Position is utilitarian, and I think he held too that the principles that meet this standard need not amount to direct utilitarianism. Direct utilitarianism of course must entail the Pareto principle, but it doesn't follow that the principles chosen as a public conception of justice will conform to the Pareto principle. My paper is about that.

JW: About primary goods? About the index of primary goods?

AG: Yes, that's what I was trying to understand in that paper. Rawls' central idea is appealing: that it's the primary goods that society is responsible for making available, and what you do with them is your responsibility. But then that does raise a big question of how it's determined what the primary goods are, how to index bundles of primary goods. You have to answer those questions to make the theory say something definite.

JW: I think what your work showed was that that's problematic. I guess that the indexing number problem has never ever been really satisfactorily dealt with.

AG: I'm not sure whether it has. I try in that paper to propose how we might get something that does the job of such an index. My more general view about Rawls is that the arguments he gives for the Difference Principle really say that the situation in the Original Position is such that the Difference Principle and what utilitarianism recommends are not substantially different. Rawls' view thus isn't really an incompatible alternative to rule utilitarianism, but amounts to saying how rule utilitarianism works out as applied to questions of economic justice.

MA: Let me just quickly follow up there. There's a famous dispute between Harsanyi's conception of the veil, which says that the veil involves equal probabilities of ending up as any person, and the Rawlsian conception, which says that it's a

decision under ignorance without probability. Rawls uses his conception of the veil and a conception of how to choose under ignorance in order to get to the Difference Principle. So, are you suggesting that Harsanyi's conception of the veil was better than Rawls'? I believe that Parfit in *On What Matters* (Parfit 2011) takes Harsanyi's side in that dispute. Parfit's view, I believe, is that if we're trying to model impartiality, the stipulation of equal probability as opposed to ignorance is a better model.

AG: It's puzzling whether ruling out probabilities the way Rawls does makes a real difference. What parties in the Rawls' Original Position are ignorant of, in Rawls' scheme, is what their society is like—beyond that the circumstances of justice obtain, including moderate scarcity. If parties knew exactly what their society was like, including the proportion of people with each set of relevant characteristics, then clearly they could use probabilistic reasoning. They would each take it that one had roughly an even chance of being female or male, and that one has a 1% chance of being in the top 1% and a 20% chance of being in the bottom 20%. So if they could specify, for each way their society might turn out to be, such things as which distribution of income was to be realized, then numbers would very much count, and they would specify whatever setup gives a person the best prospects. Parties could specify a function from what the society is like, with its economic possibilities, to what the economic institutions should be like.

But that isn't how Rawls sets things up. What the parties choose is not a standard that the economic setup is to satisfy, but a public conception of justice, a conception that the society is not only to match but to adhere to and be motivated by. The question the parties face is what the public conception of justice is to be, and this choice is made under non-probabilistic uncertainty about what their society is like and hence what the consequences of the choice of a public conception of justice will be. Rawls thought that his Difference Principle would be chosen as a part of the public conception of justice.

It's important to note also, though, that his rationale for this includes what amounts to an argument that under circumstances that the parties know to obtain, instituting the Difference Principle will maximize expected utility. The parties, he says, care little about the gains they could make above what the Difference Principle accords them, but alternative principles allow outcomes they abhor. In light of all that, an indirect utilitarian would choose the Difference Principle just as surely as would parties to the Original Position as Rawls specifies it.

JW: Is what's relevant that Rawls is stripping someone behind the veil from information which you think is relevant for making a decision?

AG: Yes, he strips them of information that would be relevant if they had it. He strips them of information as to what the economic possibilities are like that they could realize by a choice of principles to govern economic arrangements. But even more central to what he does is this: What the parties choose in the Original Position is not just the standards that economic arrangements are to meet, but what conception of justice shall be public among them. What conception of justice will be publicly accepted makes a crucial difference as to what economic arrangements will be realized and whether they will be stable.

Rawls misses an argument I just gave, but sets things up in such a way that the upshot of this argument is inconclusive. The argument he misses is this: That if in the Original Position, what's to be chosen are the standards that economic arrangements are to meet, with the assurance that whatever they choose will be implemented, then Rawls' system will amount to an indirect form of utilitarianism. They won't know what their society is like and what its economic possibilities are, but they can choose a complicated standard that, in effect, chooses for each way their society might be whatever economic arrangements would maximize people's economic prospects.

But as Rawls sets things up, they have to choose a conception of justice that is to be publicly accepted no matter what the economic possibilities turn out to be like. One candidate for playing this role goes as follows: "Give those in the worst-off starting positions the best prospects possible." That's the candidate that Rawls argues they would choose. Other candidates would be more directly utilitarian: "Realize whatever economic arrangements would give the representative person the best economic prospects." Rawls argues, in effect, that absent information about what one's society's economic possibilities are, parties to the Original Position will prefer the Difference Principle. But as I have been suggesting, how all this might depend on features of the Original Position as Rawls stipulates them is a tricky matter. And beyond that, there's the more basic question of why the Original Position should have those features. Giving it these features is supposed to yield the outcomes that Rawls wants, to be sure, but more importantly, it's supposed to illuminate the kind of rationale that ascribing these features to the Original Position might have.

MA: So is it fair to say that you strip away information from a contractor about who she is? But then the question is: For all the different possible people, does the contractor assign those equal probabilities or does she choose without any probabilities at all? Rawls seems to take the latter approach and from there he gets to the Difference Principle by virtue of the view that choosing under complete ignorance you should maximize the worst possible position. But again, if one were to strip away information about your identity, but then assign equal probabilities to the different possible societies and your being one or another person in those societies, one could then just apply expected utility theory to that choice.

AG: I'm saying that as Rawls sets things up, the prime uncertainty that the parties can't have subjective probabilities about concerns what their society is like, with its economic possibilities. If the parties knew all that, they could ascribe probabilities easily. If they knew everything there was to know about the economic circumstances of the bottom 20%, then trivially, they would have a 20% subjective probability of being in those circumstances. But they don't know such things as what the distribution of income and opportunities is for those in the bottom 20% of starting positions. So I agree with Rawls that it doesn't make any sense to ask what's the probability that this society is twenty-first-century America or ancient Greece or something else; in the Original Position, I take it I am in any of the vastly different possible societies where the circumstances of justice apply. We can agree with Rawls that it doesn't make any sense to say that people in the Original Position assign probabilities to being in certain kinds of societies.

I was also insisting that given any particular way society might be apart from who's who, the thing to choose is to maximize expected utility. So, there's a sure-thing principle that says: "Even without knowledge of what your society is like, maximize expected utility." But that's not an option that parties to the Original Position can choose. They have to choose a conception of justice to be accepted in their society, and they have to make this choice in ignorance of what their society is like in its economic possibilities.

MA: O.K, I'm going to take over asking questions at this point. Welfarism is the view that social alternatives should be evaluated solely in terms of the individual well-beings associated with them. The Difference Principle, with its focus on opportunities, is non-welfarist. In a provocative 2001 article, Kaplow and Shavell (2001) argued that any non-welfarist principle violates the Pareto principle. Faced with such a conflict, what do you think should be abandoned? Or does it depend on the non-welfarist principle that's being considered?

AG: My sympathies are with the Pareto principle and welfarism. I talked a long time ago about the "intrinsic reward" of a life (Gibbard 1986), and setting up whatever system gives the best prospects for intrinsic rewards of people's lives seems to me to be a good thing to favor. That will require satisfying the Pareto principle, couched in terms not of preferences but of the intrinsic reward of lives. But the conception of justice whose public acceptance would most foster intrinsic reward needn't satisfy the Pareto principle in any form.

JW: Faced with such conflicts, we're interested in how you would resolve the impasse of the incompatibility.

AG: It's a difference between direct utilitarianism and various forms of indirect utilitarianism. Indirect utilitarianism evaluates not acts or economies directly, but such things as possible ethoses for moral standards or standards of economic justice. Such evaluations will satisfy a kind of Pareto principle: If one ethos yields prospects that are better for someone and worse for no one, then that ethos is preferable. But the dicta that comprise the ethos don't have to include the Pareto principle and don't have to be consistent with the Pareto principle.

That said, one feature of utilitarianism that worries me is that you can try to base things on reciprocity, and then utilitarianism says that even if others aren't reciprocating, one must take their good equally into account except as taking their good into account produces incentives that aren't good-promoting. So I do have that worry about utilitarianism that it doesn't sufficiently cater to demands of reciprocity. I'm worried about Rawls' contention, which seems right, that motives of fair reciprocity are stronger than motives of unconditional altruism. Still, if you can get everybody to cooperate in a scheme, if I have my choice of what scheme everybody will cooperate in, I think we should go for one that produces the highest expected intrinsic reward of people's lives.

MA: Let me ask a different question and then we'll come back to this issue of the strains of commitment in utilitarianism. Your analysis of Rawls' Difference

Principle employs a formal economic model in order to address normative questions. In a 1978 article with Hal Varian (Gibbard and Varian 1978), you considered how economic models can be used to help explain features of the real world. You argued that the usefulness of an economic model for explanatory purposes when applied to a real-world situation is due to the assumptions being either sufficiently close to the truth or a caricature of the key features of a situation with the conclusions being robust to variations in the caricature. What relevance does this analysis of the value of economic models for positive economics have for normative purposes?

AG: Well, let me say first that that paper came from my being asked to give an American Philosophical Association paper on Philosophy of Economics, and I thought I didn't really have the subject well enough under my command to do it all by myself, and Hal agreed to collaborate on it. So, my motivation in writing it didn't have a lot to do with normative questions. But I kept finding that philosophers thought that economic models were ridiculous because they had assumptions that obviously weren't true of the world. And so I was just trying to educate philosophers about how economic theorists think of their models that economic theorists aren't the idiots that philosophers were assuming that they were. Of course, we use models a lot for normative purposes, as with the model of the Original Position or Harsanyi's model, but that wasn't what our paper was about. It was really about positive applications of economic models.

JW: In many normative analyses, simple models drawn from economics are used to work out ideas or test ideas, and they involve making unrealistic assumptions. Is this a shortcoming? Or are there some kinds of rationales analogous to what you did for positive economics that justify the use of simple models of production and distribution to, say, test a theory of distributive justice?

AG: I would think the latter. I don't have elaborately worked out views on that. But in order to scrutinize the things you might want to say about messy reality, you need to have some tractable way of thinking about matters, and thinking that messy reality has some important relation to a tractable model is about the only strategy of inquiry that is available. I would think one should always keep one's eye on what the tractable oversimplified model might have to do with reality, but—well, it's pretty much the way Arrow talks about positive models. You say, "Well if reality were this way, here's what we could conclude." And then the next question is, "Are there features of reality that make a difference and tell us we shouldn't conclude that?" So, say a normative argument such as an invisible hand argument for laissez-faire abstracts from externalities, and it also abstracts from information, the way the Arrow–Debreu model abstracts from information. If things were the way that the Arrow–Debreu model (Arrow and Debreu 1954) supposes, then laissez-faire would be Pareto efficient. That leaves a number of important things to note.

First, of course, the Pareto principle tells you nothing about distribution, and it has always seemed to me that the old arguments that utilitarianism will have an egalitarian tendency are good arguments, despite being despised from the 1930s on. And so, laissez-faire is not the thing to go for. It's true that feasible ways of trying

to increase equality will violate a version of the Pareto principle, but that's a cost to be born. And it's a version of the Pareto principle that basically talks about the ways you could organize society with perfect knowledge and unlimited calculating power. A Pareto principle that abstracts from that may have something to do with reasonable policy choice, but not a lot.

And then, of course, the story that standard economics tells says that we have ignored externalities, and we'd better bring those in. We've ignored information, and later on, people studied intensively systematic ways of bringing information in. And besides all that, there are the ways in which we can't be perfectly rational expected utility maximizers, so that nudges and things like that may come in further on in the discussion. So, I would think that the idealized models for normative purposes, as well as positive purposes, tell us what would hold if certain things were true, and the relevance of what would hold if certain things were true to what actually holds is going to be a complicated question.

MA: A different way in which this question might be relevant depends upon whether utilitarianism is simply a criterion of goodness or a decision procedure. Some people say that utilitarianism simply specifies the conditions under which an act or a possible world would be better, but is not necessarily operational as a decision procedure. On that view, models might not have a lot of direct relevance to morality. A different construal of utilitarianism says, "maximizing total well-being should be a direct decision procedure." In which case, models are very important because humans, at least, can't hold representations of whole possible worlds in their heads. Humans will need to use models of those worlds, and then we have a question whether those models are adequate. So let me ask, on your view is utilitarianism simply a criterion of goodness, or does it have more direct decisional role?

AG: Well, I think I would take the same sort of view as R. M. Hare took, that utilitarianism would be fine for what he calls archangels, but not as a full ethos for society (Hare 1980). For the latter, the question is more, "What ethos would produce the best results?" And the ethos that would produce the best results probably is not asking one to calculate each decision in a utilitarian way. As I said before, I think that the question Rawls asks amounts to this too: His question is what would most appeal to parties in the Original Position as the ethos to have as our public conception of justice.

MA: O.K. So let me ask you about causal decision theory. Leonard Savage (Savage 1954) modeled decision-making under uncertainty as a ranking of acts, where an act maps states of the world onto outcomes. For Savage, states are act independent, but this has been generalized to allow for act dependence. Few economists seem to be aware that there's an extensive philosophy literature dealing with an alternative approach to decision-making under uncertainty known as causal decision theory in which judgments of causality play an important role. Your 1978 article with William Harper (Gibbard and Harper 1978) is one of the seminal papers in this literature. For those unfamiliar with causal decision theory, please provide a brief explanation of what this theory is and why it is important. In what kinds of circumstances do you think that it is safe for economists to ignore issues of causality in decision theory?

AG: Well, first, a couple of preliminary things. One is that the Gibbard–Harper article was drawing very heavily on a discussion involving Robert Stalnaker and David Lewis (Stalnaker 1981), and we buried this acknowledgment too far in the footnotes. It was Stalnaker whose idea we were mostly developing. We thought of ourselves as writing a paper about a proposal by Richard Jeffrey (Jeffrey 1965), but in order to fill in the background, we had to explain Stalnaker's unpublished proposal. So in the last reprinting (Gibbard and Harper 1981), we started out with something like, "this paper develops a proposal by Robert Stalnaker." Second, I don't think what we're doing is an alternative to Savage. Savage requires act-independence, but then there are two kinds of independence that one might demand. One is probabilistic independence and the other is known as causal independence. I've talked to various decision theorists about this, and the people I've talked to seem to split about 50/50 about which Savage intended. I guess I tend to think that Savage must have intended what we're advocating, what came to be called causal decision theory. I would have thought that Savage wants the agent to be sure that which state obtains doesn't depend causally on what one does.

Stalnaker was responding to a theorem of David Lewis (Lewis 1976); this was at a symposium that I wasn't at, but that Harper, Lewis, and Stalnaker were, and that Harper told me and others about. Stalnaker had thought that there was a conditional if-then operator that produces a conditional proposition as a function of two propositions. So on his view, "If A, then B" is a proposition, true for some ways the world might have been and false for others. He proposed that this propositional operator had the nice feature that the probability of "If A, then B" is the corresponding conditional probability, the probability of B given A. Lewis, at that session that Harper told me about, showed that such a thing could hold in general only for excessively trivial cases. And Stalnaker said, "Well, suppose, suppose R. A. Fisher had been right when he conjectured—well, I guess Fisher just said it was a possibility for all the statistics tell us—that the correlation between smoking and cancer is not because smoking tends to cause cancer but because there's a common cause of both of them." [See Pearl (2000) for a discussion of Fisher's views.] Suppose there's a genetic factor that leads both to smoking and to cancer, so the news that I'm about to smoke becomes an indication to me that I have this genetic factor and that I'm going to have cancer. So smoking, the story is, doesn't cause cancer, but it's an indication that I'll get cancer.

One interpretation of Savage is that he requires that the states be epistemically independent of the acts—that, for instance, the epistemic probability of one's having this genetic factor be independent of whether one smokes. Such epistemic act-independence produces the kind of decision theory that says you shouldn't smoke because smoking is an indication of something you strongly and reasonably don't want. What I like to call "instrumental expected utility theory" says, in contrast, that since in this fantasy of Fisher's smoking doesn't in any way tend to cause cancer, it makes sense to go ahead and smoke if you enjoy it.

MA: So would that in fact be your view? That is, your view would be the instrumental or causal, as opposed to the evidential, approach, namely that in this scenario

where smoking doesn't cause cancer and simply indicates cancer, if you like smoking, you should smoke?

AG: Yes—and this applies to Calvinism also. Some people say that Calvinists think that because leading a godly life is an indication that one's among the elect, wanting to get to heaven is ground for leading a godly life, even at some sacrifice of earthly goods. My sister, who's a historian of that era, says, "No, the idea is that the elect really care intrinsically about godliness. So, Calvinists are said by some people to be epistemic utility maximizers but, according to my sister, they were actually instrumental expected utility maximizers.

MA: I see. But surely it's the case that economists, at least implicitly, are causal or instrumental and not evidential. In the case you are describing, I imagine economists would say that you should smoke. And in the classic case of Newcomb's paradox (Nozick 1969), I imagine economists would say, "Take both boxes, not one," even though your taking both boxes is evidence that the predictor has not put the big prize in one box. And in the case of twins playing a prisoner's dilemma, economists are surely going to say that you should rat even though that's evidence that your twin is going to rat. So, if this is right, then economists really owe you a great debt. And it also seems to me that economists, although they use the Savage framework or the Anscombe–Aumann framework (Anscombe and Aumann 1963), which is a variation, are very sloppy about ensuring that the states are causally independent from the acts.

AG: Yes, apart from not knowing about the Anscombe–Aumann framework, I'm 100% with you on all that. I do know about the Savage framework, and it, I maintain, should be read instrumentally, not epistemically.

MA: All right. You are best known to philosophers for your work on meta-ethics. Let us now turn to that subject and how it relates to your approach to utilitarianism. You characterize normative questions as planning questions, that is, questions about how one ought to act, what one ought to believe, and what one ought to feel. Please explain why you view normative "oughts" in terms of plans. What are the advantages of this perspective over competing views?

AG: Well, first, it has a disadvantage, namely that I have to use the word "plan" in a way that differs from the ordinary use in lots of ways, so . . .

JW: But it's a use that economists would be sympathetic with.

AG: So, to answer a normative question about what, say, Caesar ought to have done when he arrived at the Rubicon, I ask myself, "Well, suppose I'm in Caesar's situation." Under that hypothesis, I ask myself what to do. And if my hypothetical plan is not to cross, then that amounts to thinking that the rational thing for Caesar to have done was not to cross.

MA: Are you concerned about the problem of changing your personal identity, that it's metaphysically impossible for Allan Gibbard to be Caesar? To the extent that you want to plan not just for the things you're going to do, but for the contingency

of you being someone else, how do you get around these issues about the necessity of personal identity?

AG: Well, it's metaphysically impossible for me to be Caesar, but it's not strictly impossible epistemically for me to get conclusive evidence that I'm Caesar. It's far-fetched, but when people are mixed up in babyhood or something like that, then they have coherent beliefs about who they are that are epistemically justified but metaphysically impossible. So, it's the epistemic possibility that matters here. I mention this sort of thing in *Meaning and Normativity* (2012, p. 133) and elaborate an entire framework for such questions in Appendix 1 of that book.

JW: To follow up on that, you wrote a paper about interpersonal comparisons in the late 1980s in the Elster–Hylland volume (Gibbard 1986) where you developed in a very nuanced way some of Harsanyi's ideas about the nature of interpersonal comparisons (Harsanyi 1955). Does what you've just said tie in with your earlier arguments about the scrutability of different individuals?

AG: Well, it ties in with the question, "Suppose I'm John Weymark. What to want?" This is an intelligible question, even though it's metaphysically impossible that I be John Weymark.

MA: Let me ask you about the motivation for thinking of "oughts" or an ought statement as the expression of a plan or commitment, as opposed to an assertion of either a natural or a non-natural reality. Can you say a little more about this, again for an audience of social choice theorists who are not familiar with expressivism and the debate between expressivism and competing positions in meta-ethics?

AG: Well, I'm basically taking off from the kinds of arguments that G. E. Moore gave in around 1900 (Moore 1903) for a view that Henry Sidgwick also had (Sidgwick 1874), that no purely naturalistic assertion will have the same meaning as a statement about what I ought to do in a situation. So, if you say that in a prisoner's dilemma with one's twin one ought to rat, then there's no empirical fact for which the settling would be tantamount to deciding what one ought to do in that situation. And so that leaves a puzzle. Moore and various other people say, "Well there's a fact, but it's not a natural fact," and that seems very obscure. So, in the 1930s, people like A. J. Ayer (Ayer 1936) said, "Well, that's because what we have to explain is not what naturalistic thesis is equivalent to saying that one ought to rat, but the state of mind of believing that one ought to rat." And Ayer thought that this state of mind is emotional. But critics say, "The question isn't really what I do feel about it, but what I ought to feel about it." And it seems to me that we do ask ourselves questions about how to feel about things. "Ought I to feel jealous?" Questions like that.

And so what I tried to do was apply Ayer's strategy, not to moral assertions but to normative assertions more generally: What is the state of mind of thinking that one, say, epistemically ought to believe the theory of natural selection? What I've tried saying is that we form something like plans for what to believe given certain kinds of evidence, and what I'm doing with such a statement is expressing having such a plan for what to believe given certain evidence. I could ask, "What was it rational

for David Hume to believe about the causes of life?" Hume, I think, was basically a creationist. He thought that we could only see plants and animals as designed to do things crucial to reproduction. Before Darwin that was the rational thing to believe. Hume's most notable discussion of this is in Chap. 12 of his posthumous *Dialogues Concerning Natural Religion* (Hume 1779). How to interpret Hume's intent in that chapter, though, is immensely controversial. When I say that's the rational thing to believe in Hume's evidential circumstance, I'm forming plans for what to believe given Hume's evidence without having Darwin's proposal available.

MA: So you've just explained epistemic norms in terms of plans for belief, and in both *Wise Choices, Apt Feelings* (Gibbard 1990b) and *Thinking How to Live* (Gibbard 2003), you explain moral norms in terms of plans for feelings, that is, plans for feeling guilt or resentment. How do you respond to the objection that you can't control either your beliefs or your feelings; you can only control your decisions? And so, at most we can adopt plans for choice but not plans for belief or plans for feelings?

AG: Well, my conclusions about what it makes sense to believe do have an influence on what I actually believe. So if I am, say, running an experiment and doing a statistical analysis, statisticians always say to plan one's analysis as part of planning the experiment: Planning the analysis is planning what to believe, maybe what degrees of belief to have in various hypotheses under conditions of getting certain results in the experiment. So, you can't believe at will, but after all you can't will at will either. What I will is influenced by my beliefs about what it makes sense to will. And, likewise, what I believe is influenced by my beliefs about what it makes sense to believe.

MA: Let me come back to this issue about emotivism. In your 2006 interview with Alex Voorhoeve (Voorhoeve 2009), you mentioned that as an undergraduate at Swarthmore, you were interested in the status of morality and, quote, "disturbed by emotivism." Can you comment on the difference between expressivism and emotivism, and how in turn they relate to other positions in meta-ethics?

AG: Yes. I'm not actually sure whether the term "expressivism" stems from me or from someone else; I haven't been able to find that out. But there's something that Ayer has in common with Hare and with me, namely a strategy of explaining meanings by characterizing the states of mind that are expressed. So, Ayer explains the meanings of terms like "right" and "wrong" by saying, in effect, that to believe that something is wrong is to have negative feelings toward it (Ayer 1936). People call it the "Yay, Boo Theory," which is meant to deride the theory, but is actually a good way of thinking about its logic. R. M. Hare says that it's not a state of feeling that one's expressing with a moral assertion, but one's preferences for the case of being each person involved seriatum, or better, one's preferences for the case of being me and the case of being you and the case of being John, and the case of being each other person (Hare 1980). That is, if it's really a moral belief, then the conditional preference isn't affected by who's who in the situation. My preference for the case of being you is the same as my preference for the case of being me. Otherwise, my preferences don't constitute a moral conviction. My own view that the state of

mind of making a normative judgment "is something like having a plan" takes that same form. So I take the idea of expressivism to be that you explain meanings by explaining the states of mind involved in a way that doesn't involve helping yourself to describing them as the state of mind of believing such-and-such. The philosopher Jason Stanley said to me, "Well, isn't that just what philosophers call functionalism in the philosophy of mind?" And I think he's right.

MA: Let me ask you another question about meta-ethics. Much of meta-ethics has been preoccupied with the so-called Frege–Geach problem (Geach 1964). Can you explain for an audience of social choice theorists not expert in the topic what the Frege–Geach problem is; why many ethicists have found it so difficult; and why your approach is well positioned to address it?

AG: To answer that, it's easiest to think of straightforward emotivism of the Ayer kind, so that saying "Lying is wrong" amounts to saying "Boo for lying." But then Peter Geach says, attributing this to Frege with some justification, that we've got to explain not only sentences like "Lying is wrong," which amounts to "Boo for lying," expressing feelings against lying, but also how a father can argue, "If lying is wrong, then getting your little brother to lie is wrong too. And lying is wrong, therefore, getting your little brother to lie is wrong."

So first, we have to explain a sentence like "If lying is wrong, then getting your little brother to lie is wrong too." And furthermore, we have to explain it in such a way that the inference is valid. We don't transfer between one interpretation of "Lying is wrong" when it's self-standing and a different interpretation when it's part of the conditional, "If lying is wrong, getting your little brother to lie is wrong too." Otherwise, the inference won't be valid; it will be in the form of, "If A, then B; A^*; therefore, B," which isn't a valid inference.

So what I say is that a statement determines a set of states of mind you might be in. A state of mind would include one's factual beliefs and also one's feelings or preferences. So, what the assertion "If lying is wrong, then getting your little brother to lie is wrong" does is exclude states of mind in which you have feelings against lying, but don't have feelings against getting your little brother to lie. If you think of it that way, then all sorts of standard truth-conditional logic just comes out with a different kind of explanation for its validity, and the explanation for the validity applies to prosaically factual beliefs as well as to beliefs that amount to feelings about something.

MA: Can you say a little bit more about the role of plans in rational choice? One version of expected utility theory says that the focus at any point in time should be the choices available to you at that point in time and your information about consequences going forward. It's not rational to stick to a prior plan if a different choice maximizes expected utility going forward. Do you view planning as being more fundamental to rational choice than on this traditional version of expected utility theory that focuses on acts as opposed to plans?

AG: Philosopher Michael Bratman's main lifelong work has been on why plans are important for leading one's life, and individual decisions in isolation won't do the

job (Bratman 2014). So he's very interested in when one ought to reopen a question that one had already settled in one's planning. And I guess I would think that we need a kind of indirect approach here, asking what kind of propensity to reopen questions is optimal.

JW: To clarify a bit for economists, when you are talking about a "plan," you are thinking about what we would commonly call "contingent plans."

AG: Yes, right.

JW: You might find yourself either actually or hypothetically in some situation and the question is: What should you do in that circumstance? Is that correct?

AG: Right.

MA: Just a couple of more questions about meta-ethics. Your meta-ethical views emphasize the normative role of consistency. For example, in a striking passage from the beginning of your book, *Thinking How to Live* (Gibbard 2003), you suggest that inconsistency is a kind of normative mistake. Let me quote:

> Pluto, imagine, betrays his dear friend Minerva to get rich, leaving her impoverished and building a fancy house with the proceeds. Then in a fit of remorse and self-disgust, Pluto burns down the house, even cancelling the insurance so that his renunciation will be genuine. It's hard to claim that throughout this whole affair he acted without mistake. If it was fine to burn down the house, then it wasn't all right to betray Minerva in the first place. (p. 17)

Moreover, you require that preferences and beliefs satisfy the familiar consistency conditions from decision theory, what you call coherence conditions.

Can you explain the normative status of coherence? Is the normative statement that plans should be coherent itself just an expression of a plan, or does a requirement of coherence have a deeper, perhaps a natural, basis?

AG: Well, I would think vaguely that planning has a basis and that incoherent plans are ones such that if one aspect of them has a basis, then another aspect of them undermines that basis. That's vague, but that's the vague idea. Even though Peter Hammond's treatment (Hammond 1988) of what he calls consequentialism is a little too formidable for me to master entirely, I think the kind of analysis he gives is the kind that that's most fundamental.

MA: Take the statement, "You ought to have consistent plans." One view might be that this is simply another "ought" statement, and like any other "ought" statement, it doesn't describe a fact; it's simply my expression of some kind of very general plan to have consistent plans. And yet in your work, consistency seems to have a deeper status than that.

AG: Perhaps you don't need to have a plan at all, and so the dictum should be "Don't plan incoherently." This has a basis, namely that inconsistent plans are going to be in some respects self-defeating. Whatever might be to be said for such a plan, some alternative plan is guaranteed to do better. So a plan to eschew inconsistent plans is based on a logical feature.

MA: Intuitions also play an important role in your meta-ethics. Why do you think that coherent or consistent planning is not possible without relying on some intuitions?

AG: Well, how do we choose among different consistent contingency plans? It seems that we have no way of doing it without relying on judgment, and so some reliance on one's judgment seems inescapable. It has to be a critical reliance on one's judgment, but I don't see the alternative. Now it may be that it would be better to shape myself to be more spontaneous and less careful and not try to get my judgments to line up with each other. There are questions about what sort of person it has the best prospects for me to be. But even if I ask that, I have to rely on my judgment, and maybe on the judgment of the people I think about the problem with, to come up with an answer. Without some reliance on one's judgment, one doesn't have any basis for anything.

MA: All right. Let me ask one final question here about meta-ethics and then we can move on to questions about utilitarianism and individual goodness. There is of course a naturalist line in meta-ethics which looks to fully informed preferences. So Michael Smith's view, I take it, is that normative facts are facts about people's convergent fully informed preferences (Smith 2004), and I take it that Peter Railton has suggested a similar line both for facts about one person's good and for facts about moral good (Railton 2003). Fully informed preferences do play a role in your meta-ethics: I might plan to rely upon what I believe to be people's fully informed preferences. And yet you resist saying that "oughts" are simply assertions of fact, where the fact turns out to be a fact about fully informed preferences. Why can't we simply say that that's what "ought" facts are?

AG: I do think that there will be an ideal circumstance for forming preferences. But the question of what the ideal circumstances are is really a question of which preferences to trust. The answer that they must be fully informed might be too simplistic. The example I use in *Wise Choices, Apt Feelings* (Gibbard 1990b) is: What if I'm convinced that if I really were fully informed about what went on in people's innards, I would never want to eat with anyone else. Well, it seems that the thing to do there is enjoy other people's company when I'm eating, and ignore what's going on in their innards. So, it will have to be the right sort of being fully informed, and so there should be an account of what kind of information state is the ideal one for forming preferences.

And then we can ask: "What does that question mean?" And I say: "We can understand that as a question of what kinds of preferences to trust." If I know that something would be my preference under condition A and then something different would be my preference under condition B, shall I defer to the preference I would have under condition A or the preference I would have under condition B. This is a normative question; it amounts to asking what are the conditions such that I ought to adopt the preferences that I know I would have under those conditions.

JW: As I understand it, that's true not just for moral issues involving other people, but also in your prudential decisions. You have to trust some of your previous judgments, . . .

AG: Right.

JW: . . . your preferences; otherwise, you just can't get on with life.

AG: Right. In conversation, Brandt talked sometimes about choosing between the job that would be most prestigious and the job that one would be happiest in. And that's maybe mostly a prudential question, which we can put in the form of: "What would I prefer under ideal conditions." But regarding conditions as ideal is something we have to explain. And I propose we explain regarding them as ideal as deferring to the preferences that I learn I would have under those conditions.

MA: But there is a competing, descriptive analysis of the claim that I ought to do something. The descriptive analysis is, "I ought to do something given my current informational state, which might not be perfect, if everyone with full information, taking account of my current informational state, would want me to do that." And on this descriptive view, facts about what these fully informed advisors would want is a kind of descriptive fact. So, I view that as being a competitor to your expressivism.

AG: I don't think it is actually a competitor; part of it combines well with expressivism. I do think there will be a naturalistically formulable condition that's such that I ought to do a thing if and only if the thing meets that condition. But then we can dispute about what the condition is. This issue isn't entirely straightforward, as the example of having full knowledge of what's going on in people's innards suggests. So that still leaves the question: "Suppose people disagree about what the proper conditions for forming preferences are. What are they disagreeing about?" And I propose, they are really disagreeing about which sorts of preferences to trust, which conditions are such that the fact that preferences are formed in those conditions supports trusting those preferences or adopting them.

MA: Let's come back to themes of utilitarianism and the strains of commitment. You regard morality as being concerned with identifying how to live together on terms of mutual respect that nobody could reasonably reject. You argue that a contractarianism based on such terms would result in decisions being taken that utilitarianism recommends provided that the parties can be counted on to honor their agreements even in the face of strong personal countervailing motives (Gibbard 2008). You assume that there's full compliance. Critics of the veil of ignorance arguments of Harsanyi and Rawls have argued that in the non-ideal world in which we live, individuals would not necessarily abide by the agreements reached behind the veil once it is lifted. In Rawls' terminology, the "strains of commitment" are too great. How much confidence do you have that your utilitarian conclusions would survive if full compliance were not assumed? Relatedly, what relevance does ideal world theorizing have for thinking about morality in non-ideal circumstances?

AG: I think of something both in the spirit of Brandt (1979) and in the spirit of Rawls (1971) as applying here. We can imagine choosing a social ethos. Brandt

talks about what moral code to support for one's society, and then has a complex description of what a moral code is. Rawls has phrases that amount to the same thing; he talks of a "public conception of justice." So in the Original Position, one chooses among alternative public conceptions of justice that we might have and institutionalize. Having something as the moral code of one's society, its public conception of justice, doesn't by itself produce full compliance. It has an influence, but not always an overriding influence. So what's to be chosen from behind the veil of ignorance will depend on what people will actually do if one set of standard or another comprises the public conception of justice. And some possible public conceptions of justice are going to elicit more compliance than others.

I think that despite the fact that Rawls seemed to deny it, his view amounts to a kind of indirect utilitarianism applied to broad features he attributed to the social world. Actually, my recollection is that on close reading, we see that Rawls avoided addressing indirect utilitarianism, and just denied that the parties behind the veil of ignorance would choose direct utilitarianism as the public conception of justice. That's quite consistent with the kind of rule utilitarianism that Brandt advocates. I thus claim that Brandt and Rawls are advocating the same sort of conception of justice, Brandt perhaps more straightforwardly and Rawls perhaps more convincingly.

MA: I have a question about a person's "good." You identify a person's "good" with whatever plays this role when we plan to live with each other on the terms described earlier. The concept of "personal goodness" is given content by specifying three axioms that this plan for living with others should satisfy. In your own words (Gibbard 2008), these axioms are as follows:

> (1) "Prefer most to live with others on a basis that no one could reasonably reject on his own behalf." (p. 51)
>
> (2) "A rejection on one's own behalf of a going social arrangement is unreasonable if, absent information about which person one would turn out to be, one would have rationally chosen that arrangement on one's own behalf." (p. 51)
>
> (3) "One chooses rationally on one's own behalf only if one chooses what is prospectively most to one's good." (p. 52)

Why do you think that these requirements lead to a determinate quantitative measure of personal good? On what basis are these personal goods interpersonally comparable? Truistic but vague notions about my own well-being—what's good for me, what a good life would be, choosing on my own behalf—are all different ways to talk about my "goodness." But, do those notions, even for a particular person, have determinate content?

AG: I don't have a clear memory of what I was doing in setting down this account, but I'll try to rethink. What's to a person's good, on the kind of account I was advocating, is a normative question, a question of what to have play a certain kind of role in a conception of social justice. So questions of determinacy and interpersonal comparability are aspects of the broad normative question of how to live together on a basis of mutual respect. I'm trying to adopt the Rawlsian (Rawls 1971) and Scanlonian (Scanlon 1998) strategy of thinking about what no one could reasonably

reject, by standards that rule out, among other things, its being the case that for every possible standard of justice there will be someone who could reasonably reject it. These specifications appeal to the notion of rejecting a standard "on one's own behalf," and don't by themselves specify what this amounts to. The idea is that one can't give determinate content to talk of "your good" or "my good" in advance of facing the ethical question of how good is to be distributed among us. In saying this, I think I am agreeing with Scanlon.

As for how determinate the notions of your good or my good are, these are then normative questions; they boil down to questions of what to choose in various hypothetical situations. We have to make actual choices, and if we are capable of thinking deeply about being in various circumstances, then we can make hypothetical choices. So the content of notions like "my good" is determined by the answers to questions that we answer by making hypothetical choices, or forming hypothetical preferences. That said, we are left with the broad question of how to live with each other on a basis of mutual respect and what standards govern this.

As for my own views on this, I'd say this: If, as I hope, we can make sense of the idea of "the intrinsic reward of ones life," I'd want that to figure in what rejecting a standard "on one's own behalf" amounts to. I think, though, that in the Tanner Lectures (Gibbard 2008), I wanted to be agnostic about much of this.

JW: In Harsanyi's Impartial Observer Theorem (Harsanyi 1955), the objective is to socially rank uncertain prospects from a moral point of view. He imagined an impartial observer being placed behind a veil of ignorance believing that he has an equal chance of being anybody once the veil is lifted. Assuming that his preferences behind the veil satisfy the expected utility axioms, Harsanyi argued that the impartial observer would rank alternatives according to their average utility. Now, you don't use this theorem in exactly the same form as Harsanyi; you talk about this in terms of "personal goodness." Can you explain what's different about your use of Harsanyi's thought experiment and his own, and why that's important?

AG: First—though I think Harsanyi would agree with me on this—one can't take a person's preferences all told as an indicator of that person's good. One's preferences are influenced not only by one's judgments of one's own good, but by one's judgments of the good of others. And so I struggle trying to see what the notion of a person's good can be. It's a notion that Scanlon (Scanlon 1998) doesn't think can be made coherent in advance of asking how a person's good is to figure in our conception of social justice.

JW: It seems then that your concept of personal goodness is designed so that you can take these considerations into account in a way that is not so transparent at least in Harsanyi.

MA: One quick follow up: As you well know, there's a big debate among preference theories of well-being about how to restrict preferences to be self-regarding.

AG: Yes.

MA: Brandt doesn't really deal with this in his book (Brandt 1979). And then Mark Overvold has a series of papers where he tries to explain self-regarding preferences

in terms of preferences that are existence entailing (e.g., Overvold 1980), and that in turn has been criticized. So, are you saying that you want to circumvent that whole debate by thinking of personal good, as opposed to preferences either restricted or not? Are you saying that specifying personal good as playing these various roles is the solution to what may be an insoluble problem for preference-based views?

AG: Right. I think the question of what's to a person's good isn't strictly empirical. It is, as I have been saying, a question of what's to play the role of a person's good in our view of how to live together.

MA: Right.

AG: That was a question when we had our regular once or twice a week lunch with Richard Brandt and William Frankena for the first couple of decades after I came to Michigan. What a person's good consists in was something that we discussed quite a bit, and it never seemed to me that there was a simple, straightforward answer independent of what sort of role individuals' goods are to play in our general ethical theory.

MA: So, let me come back to utilitarianism squarely. One prominent critique of utilitarianism has been that it ignores the distribution of well-being. So-called prioritarians advocate a generalized form of utilitarianism, whereby well-being numbers are transformed by a concave transformation and then summed, which has the effect of giving greater weight to well-being changes affecting worse-off individuals. A second critique of utilitarianism has been that it ignores considerations of individual desert and responsibility. How might the utilitarian respond to these critiques?

AG: Well, on the second question, I would think the kind of answer Rawls (Rawls 1971) gives is right, that an ethos of desert and responsibility we can think of as being part of social mechanisms for promoting the goals to be promoted. So, we don't come to questions about how to arrange our social life already equipped with answers as to who deserves what. Our question is what to treat as desert in our social thinking and feelings—especially when we distinguish, as Rawls does, between desert and entitlement. Whether a team deserved to win is a different question from whether the team was entitled to the status of being the winner. And so to study desert we would have to ask, "What is the role of thought and feelings about desert, as opposed to entitlement?" And then, "How shall we arrange an ethos of desert to serve the purposes that it makes sense to want served socially?"

MA: I see. So on this view desert is a feature of social practices or understandings, and the question is which such practice maximizes utility. But there's a line in moral philosophy starting with Ronald Dworkin (Dworkin 2000), including people like Jerry Cohen (Cohen 2011) and Richard Arneson (Arneson 2011), which says that desert or responsibility has a much more bedrock role. Now, they're egalitarians, not utilitarians, but the thought would be that at the bedrock, if two people are badly off and we have to choose between giving a well-being improvement to the one who's badly off, as a matter of option luck, from bad choices, or giving it to the one who is badly off because of brute luck, we should do the latter—that as a kind of a bedrock

matter of moral rightness, we should take into account responsibility or desert. So, I take it that you are resistant to that thought.

AG: Yes, I would think we have to develop an ethos of responsibility that promotes the intrinsic reward of people's lives, and part of what's going to come out of that is systems of incentives, including incentives of being regarded as deserving.

Please remind me what the other part of your question was.

JW: About prioritarianism.

MA: Right. So the second and orthogonal critique of utilitarianism comes from this idea of prioritarianism, which Parfit (Parfit 2000) originates and has been around now for at least twenty years. It says that well-being itself has diminishing marginal ethical significance. For those who are at lower well-being levels, giving them a given well-being improvement has greater ethical weight. It's a kind of welfarism, but it's not utilitarianism, which says that well-being changes have the same ethical weight regardless of the well-being level of the person receiving them. How would you respond to the prioritarian critique that utilitarianism is not sufficiently concerned about distribution? Utilitarians are concerned about the distribution of income, but they're not concerned about the distribution of well-being itself.

AG: Well, that supposes that we have a cardinal conception of well-being that is independent of what's to be preferred given difficult choices, choices under uncertainty where the intrinsic reward of one person's life may be maximized one way and the intrinsic reward of another person's life maximized a different way. Harsanyi didn't use the term "utility"' that way (Harsanyi 1955). He thought of a utility scale as indicating the choices that people are disposed to make, so that if a person stresses differences in how badly off she will be in a bad eventuality, the differences she stresses in her choices count as big differences in "utility."

I want to put this all not in terms of how a person is in fact disposed to choose but how it is rational for her to choose. Suppose, then, it is rational to be indifferent between a sure income of $20-thousand a year and an even-chance lottery of $10-thousand a year and $40-thousand a year. Then, the utility difference between $10- and $20-thousand a year will count as the same as the utility difference between $20- and $40-thousand a year. The utilitarian will be prioritarian in terms of income, but it could make no sense to be prioritarian in terms of "utility" in this sense. There's no such thing as the declining marginal utility of utility. It might still make sense to be prioritarian in terms of intrinsic reward, if some cardinal notion of intrinsic reward is sensible.

Whether we should be prioritarian in terms of intrinsic reward is another matter. On the direction I want to take Harsanyi's way of thinking, this depends on whether it is prudentially rational to be prioritarian in terms of one's own intrinsic reward. We might conclude that the declining marginal utility of income stems entirely from declining marginal intrinsic reward we derive from income.

So I take what is roughly the old-fashioned view of people like Arrow (Arrow 1977) and Savage (Savage 1954) that utility is a matter of what to prefer among uncertain prospects. It thus becomes analytic that expected utility is what is to be

maximized. Utility is just an index that comes out of rational choices. Sen (2002) always seemed to be a native speaker of utility language in a way that I don't understand. I think utility just is a matter of how to choose given choices about uncertain prospects. In effect, the ethical claim is that the weightings that are prudentially rational for individuals are the one's society should adopt.

MA: We've already talked about Harsany's Impartial Observer Theorem. Let's talk about his Social Aggregation Theorem (Harsanyi 1955). In the Tanner Lectures (Gibbard 2008), you use a version of Harsanyi's Social Aggregation Theorem to justify a weighted form of utilitarianism. You argue that any outcome that is not Pareto optimal would be reasonably rejected by an ideal social contract. Individuals evaluate the goodness of outcomes on a personal goal-scale. Provided that the set of feasible combinations of goodness values is convex, any Pareto optimal outcome can be obtained by maximizing a weighted sum of the individuals' goods. This is a form of weighted utilitarianism that provides what you call a "coherent common rationale" or "common goal-scale" for us to adopt. Does your theory provide any constraint on what the weights used to aggregate the individual goods must be?

AG: First, I should mention, John Broome (Broome 2008) says that it isn't really Harsanyi's Theorem I'm using, but a theorem that has been long known and that, in a well-chosen phrase, he calls "the Tangent Theorem." He thinks, however, that the Tangent Theorem doesn't establish what I claim it does.

Now, clearly the argument I gave—my argument for the particular conclusion that I was claiming follows from what I was saying—that argument doesn't, all by itself, establish anything about what the weightings should be. Its real force, if I'm right, is to rule out a feature that the preponderance of recent philosophical views of justice share. Obviously, the question of what the weights should be is very important, maybe of central importance, but this particular result doesn't answer that question.

The weights, moreover, are by far not the only thing that this argument of mine, if it is correct, leaves unsettled. I don't start out making any assumption concerning what a person's good is, or even whether the notion of a person's good makes real sense. I do suppose that we can make sense of Scanlon's talk of a person's reasonably rejecting a proposed social contract (Scanlon 1998). I suppose that the upshot will be a social contract that no one could reasonably reject, and that people will act rationally within its constraints—where acting rationally includes satisfying the standard formal requirements of coherence in policies for action. The contract, then, will reflect the bases on which individuals might reasonably reject alternative possible contracts in favor of that one, whatever those bases might be.

The argument refutes the possibility that with the contract in force, different people would be pursuing somewhat conflicting goals. For if this were so, then a possible alternative contract would be available that served everyone's aims better, and it would have been reasonable to reject the contract that was in force in favor of this alternative that serves everyone's goals better. I'm not, then, saying that a satisfactory treatment of social justice stops at the point I reached, just that this is a point that we can get to by the kind of argument I was giving.

MA: Two further questions: If the personal goal-scale is not expectational, the feasible set may not be convex. Do your utilitarian conclusions depend on convexity? Also, you restrict attention to a single feasible set, but if planning is all about considering contingencies, then all possible feasible sets need to be considered. In your response to John Broome in the Tanner Lectures (Gibbard 2008), you offered some thoughts on this issue. Can you briefly tell us what they are and if you would revise them in any way now?

AG: The argument I gave for a common goal-scale does depend on convexity, and so that leaves the question of whether its conclusions bear on cases of non-convexity. This question raises complications that I don't claim to have worked through, but if we should all work for the same basic ends in a case where the set of possibilities is convex, then I find it hard to see why non-convexity should make a difference.

MA: You earlier said that having established weighted utilitarianism, it's a further question what the weights are. But at that point, assuming one's established weighted utilitarianism, why not just adopt an axiom of impartiality or anonymity to argue that the weights should be equal? Why should there be a serious ethical question at that point?

AG: I, myself, very much agree. But I'm also asking whether logical results that we can establish can force others to agree with you and me on this issue. David Gauthier has maintained that the social contract should efficiently realize the advantages of those who would be in a strong bargaining position in negotiating the social contract, and one's bargaining position depends on which alternatives are feasible (Gauthier 1986). I think there's a way to deal with his contentions, saying that we form the social contract from a standpoint before it's been settled who has the features that confer bargaining advantages. But that may require arguments that go beyond what I gave in my Tanner Lectures.

As for the feasible set not being fixed, I argue that the social contract must be agreed to, hypothetically, in advance of extensive information as to what's feasible and what isn't. Otherwise, it isn't much of a contract; it isn't much of a contract if people insist on renegotiating when news arrives bearing on what's at stake for them in an issue. So what parties face, as they reasonably reject various proposed contracts or not, is a set of prospects. Information keeps arising concerning how the social contract will bear on one's prospects and choices, but the contract restricts how to respond to such news. In making choices as to how to respond to this new information, one is to apply a fixed goal-scale and choose the prospects that rank highest on that goal-scale.

MA: Let me ask a related, but different question. You suggest in the Tanner Lectures that the ideal social contract will require the individuals to adopt a common goal-scale as their own goals. You note that if individuals pursue separate goals, prisoner's dilemma situations may arise where an outcome is suboptimal in light of all the goals. In short, your position seems to be not only that weighted utilitarianism is correct from the ethical standpoint, but that individuals are rationally required to adopt the ethical standpoint as their own. A different tradition in utilitarian thought,

going back to Sidgwick (Sidgwick 1874), suggests that it is rationally permissible for individuals to pursue their own interests. Indeed, this seems to be Harsanyi's view as well (Harsanyi 1955). He uses the Impartial Observer and Aggregation Theorems to specify the content of ethical—that is, impartial—preferences, but Harsanyi does not argue that individuals are required to act on ethical or impartial preferences. How do you see your difference from Harsanyi in this regard?

AG: There's no such difference between me and Harsanyi, as far as I can see: It's one question what ethics requires, and another question whether rationality, in the restricted sense of coherence in one's policies for action, requires ethics. I don't have a proof that ideal coherence in action requires impartiality, and I don't think that such a proof can be given. The claim that coherence by itself requires ethics is false. That's something that I argued in my 1999 review essay on Christine Korsgaard's book, *The Sources of Normativity* (Gibbard 1999).

The ethical question, I am taking it from Rawls (1971) and Scanlon (1998), is what social arrangements and ethos would it be unreasonable for anyone to reject, given that everyone has the goal of establishing an ethos that it would be unreasonable for anyone to reject. Scanlon has this formulated much better than I'm managing to, and so take his formulation. We all have reason to prefer a social ethos where we benefit from others. We just do better if we aren't just producing an enormous prisoner's dilemma socially. So everyone has reason to prefer that social ethos given the choice. But that doesn't establish that it would be irrational to free-ride on others' adherence to the social contract, if one can.

MA: Aren't you worried not only about the strains of commitment, but the critique, which goes back to people like Shelly Kagan (Kagan 1989) and Sam Scheffler (Scheffler 1982), that says that utilitarianism as a personal standard is just incredibly demanding. That it's just not feasible in light of at least current psychology for people to actually adopt "maximizing overall well-being" without any preference for their own interests as their actual, day-to-day goal-scale.

AG: Well, the standpoint that weighs everybody equally into the social goal-scale I think of as a standard for comparing possible social ethoses. And a question about any possible social ethos is to what degree would having that as one's society's ethos actually elicit kinds of behavior that is rational for us to want from each other. So I would think these questions about demandingness and the strains of commitment and so forth come at the stage of trying to choose a social ethos. A social ethos that didn't have much influence on what anyone actually did toward others isn't one to want as a social contract. It would be a dead letter.

MA: I take it then that the common goal-scale, namely to maximize, you know, total well-being, comes in as you say at the level of a social ethos or social structure and not necessarily at the level of day-to-day choices. So, is that similar to Rawls' notion?

AG: I think so. Rawls (1971) sets up the question of what parties to the Original Position will want for their society. They then choose standards of justice to be

institutionalized and publicly recognized—a social ethos, as I am putting it, or a moral code for the society, as Brandt (1979) puts it. So on any such account, the strains of commitment will bear heavily on the choice of a social ethos, the choice of a moral code for one's society.

MA: Is that in effect parallel to Rawls' notion that the principles of justice apply to the basic structure of society, but not the day-to-day choices?

AG: Yes, very much so. In preparing the published version of my Tanner Lectures, *Reconciling Our Aims* (Gibbard 2008), and cutting my reply down to size, I somehow unwisely persuaded myself to cut a passage stressing that I wasn't advocating direct utilitarianism, that the common goal-scale wasn't directly to guide our choices of what to do. That was a bad decision; reviews of the book mistakenly took me to be advocating direct utilitarianism.

In this regard, as you indicate, I'm very close to Rawls. Rawls has the Original Position as giving the standard for evaluating basic structures. The common goal-scale, I'm saying, applies to that, as what would guide people in the Original Position in choices of basic structures. The upshot is a kind of indirect utilitarianism, and I claim that what Rawls' arguments support really is a kind of utilitarianism. I should reiterate that what goes into the common goal-scale isn't decided by the argument I gave, even if the argument is successful. It's whatever is relevant to reasonable rejection, what a person might reasonably appeal to in order to reject an ethos and its implementation. The argument doesn't show this to be welfare, unless we define a person's "welfare" as whatever enters into the common goal-scale with respect to that person.

MA: I see.

JW: Allan, in concluding this conversation, Matt and I want to say how much we appreciate you taking the time to talk to us today.

AG: Well, I immensely appreciate not only you two taking the time and you, John, taking this whole trip, but especially the wonderful, thoughtful questions you put to me on the basis of such careful reading, such careful research and reflection. I'm sorry I haven't been on top of everything, since these are mostly things I did a long time ago, but it has been a delight to try to rethink matters with the two of you. Thank you both.

Acknowledgements We are grateful to Briana Brake and Susan Hinson from the Duke Law School for their assistance in preparing a transcript of this interview.

References

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, *34*, 199–205.

Arneson, R. J. (2011). Luck egalitarianism–A primer. In C. Knight & Z. Stemplowska (Eds.), *Responsibility and distributive justice* (pp. 24–50). Oxford: Oxford University Press.

Arrow, K. J. (1951). *Social choice and individual values*. New York: Wiley.
Arrow, K. J. (1977). *Essays in the theory of risk-bearing*. Chicago: Markham.
Arrow, K. J., & Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*, *22*, 265–290.
Ayer, A. J. (1936). *Language, truth and logic*. London: Gollancz.
Brandt, R. B. (1963). Toward a credible form of utilitarianism. In H.-N. Castañeda & G. Nakhnikian (Eds.), *Morality and the language of conduct* (pp. 107–143). Detroit: Wayne State University Press.
Brandt, R. B. (1979). *A theory of the good and the right*. Oxford: Oxford University Press.
Bratman, M. E. (2014). *Shared agency: A planning theory of acting together*. Oxford: Oxford University Press.
Broome, J. (2004). *Weighing lives*. Oxford: Oxford University Press.
Broome, J. (2008). Comments on Allan Gibbard's Tanner Lectures. In Gibbard 2008 (pp. 102–119).
Cohen, G. A. (2011). *On the currency of egalitarian justice, and other essays in political philosophy*. Princeton, NJ: Princeton University Press [Edited by M. Otsuka.].
Dworkin, R. (2000). *Sovereign virtue: The theory and practice of equality*. Cambridge, MA: Harvard University Press.
Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.
Geach, P. T. (1964). Assertion. *Philosophical Review*, *74*, 449–465.
Gibbard, A. (1965). Rule-utilitarianism: Merely an illusory alternative? *Australasian Journal of Philosophy*, *43*, 211–220.
Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, *41*, 587–601.
Gibbard, A. (1974). A Pareto-consistent libertarian claim. *Journal of Economic Theory*, *7*, 388–410.
Gibbard, A. (1977). Manipulation of schemes that mix voting with chance. *Econometrica*, *45*, 665–681.
Gibbard, A. (1979). Disparate goods and Rawls' difference principle: A social choice theoretic treatment. *Theory and Decision*, *11*, 267–288.
Gibbard, A. (1986). Interpersonal comparisons: Preference, good and the intrinsic reward of a life. In J. Elster & A. Hylland (Eds.), *Foundations of social choice theory* (pp. 165–193). Cambridge: Cambridge University Press.
Gibbard, A. (1990a). *Utilitarianism and coordination*. New York: Garland. [Originally submitted as a Harvard University Ph.D. thesis in 1971,]
Gibbard, A. (1990b). *Wise choices, apt feelings: A theory of normative judgment*. Oxford: Oxford University Press.
Gibbard, A. (1999). Morality as consistency in living: Korsgaard's Kantian lectures. *Ethics*, *110*, 140–164.
Gibbard, A. (2003). *Thinking how to live*. Cambridge, MA: Harvard University Press.
Gibbard, A. (2008). *Reconciling our aims: In search of bases for ethics*. New York: Oxford University Press [Edited by B. Stroud with commentaries by M. Bratman, J. Broome, & F. M. Kamm.].
Gibbard, A. (2012). *Meaning and normativity*. Oxford: Oxford University Press.
Gibbard, A. (2014a). Intransitive social indifference and the Arrow dilemma. *Review of Economic Design,* **18**, 3–10. [Originally written in 1969–1970.]
Gibbard, A. (2014b). Social choice and the Arrow conditions. *Economics and Philosophy,* **30**, 269–284. [Term paper for the Arrow–Rawls–Sen 1968–1969 Harvard seminar.]
Gibbard, A., & Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In C. A. Hooker, J. J. Leach, & E. F. McClellen (Eds.), *Foundations and applications of decision theory. Volume I: Theoretical foundations* (pp. 125–162). Dordrecht: D. Reidel.
Gibbard, A., & Harper, W. L. (1981). Counterfactuals and two kinds of expected utility. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance and time* (pp. 153–190). Dordrecht: D. Reidel.
Gibbard, A., & Varian, H. R. (1978). Economic models. *Journal of Philosophy*, *75*, 664–677.
Guha, A. S. (1972). Neutrality, monotonicity, and the right of veto. *Econometrica*, *40*, 821–826.

Hammond, P. J. (1988). Consequentialist foundations for expected utility. *Theory and Decision*, *25*, 25–78.
Hare, R. M. (1980). *Moral thinking: Its level, methods and point*. Oxford: Oxford University Press.
Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, *63*, 309–321.
Hodgson, D. H. (1967). *Consequences of utilitarianism: A study in normative ethics and legal theory*. Oxford: Oxford University Press.
Hume, D. (1779). *Dialogues concerning natural religion*. London.
Jeffrey, R. C. (1965). *The logic of decision*. New York: McGraw-Hill.
Kagan, S. (1989). *The limits of morality*. Oxford: Oxford University Press.
Kaplow, L., & Shavell, S. (2001). Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy*, *109*, 281–286.
Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.
Lewis, D. K. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*, 297–315.
Lyons, D. (1965). *Forms and limits of utilitarianism*. Oxford: Oxford University Press.
Mas-Colell, A., & Sonnenschein, H. (1972). General possibility theorems for group decisions. *Review of Economic Studies*, *39*, 185–192.
Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.
Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday* (pp. 114–146). Dordrecht: D. Reidel.
Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.
Overvold, M. C. (1980). Self-interest and the concept of self-sacrifice. *Canadian Journal of Philosophy*, *10*, 105–118.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Parfit, D. (2000). Equality or priority? In M. Clayton & A. Williams (Eds.), *The ideal of equality* (pp. 81–125). London: Palgrave. [Delivered as the Lindley Lecture at the University of Kansas in 1991.]
Parfit, D. (2011). *On what matters* (Vol. 1). Oxford: Oxford University Press.
Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.
Putnam, H. (1975). The meaning of "meaning". *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.
Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*, 20–43.
Railton, P. (2003). *Facts, values, and norms: Essays toward a morality of consequence*. Cambridge: Cambridge University Press.
Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
Regan, D. H. (1980). *Utilitarianism and co-operation*. Oxford: Oxford University Press.
Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, *10*, 187–217.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.
Scheffler, S. (1982). *The rejection of consequentialism: A philosophical investigation of the considerations underlying rival moral conceptions*. Oxford: Oxford University Press.
Schick, F. (1969). Arrow's proof and the logic of preference. *Philosophy of Science*, *36*, 127–144.
Sen, A. K. (1970a). *Collective choice and social welfare*. San Francisco: Holden-Day.
Sen, A. K. (1970b). The impossibility of a Paretian liberal. *Journal of Political Economy*, *78*, 152–157.
Sen, A. K. (2002). *Rationality and freedom*. Cambridge, MA: Harvard University Press.
Sidgwick, H. (1874). *The methods of ethics*. London: Macmillan.

Smith, M. (2004). *Ethics and the a priori: Selected essays on moral psychology and meta-ethics*. Cambridge: Cambridge University Press.
Stalnaker, R. (1981). Letter to David Lewis of May, 21, 1972. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance and time* (pp. 153–190). Dordrecht: D. Reidel.
Temkin, L. S. (2012). *Rethinking the good: moral ideals and the nature of practical reasoning*. New York: Oxford University Press.
Voorhoeve, A. (2009). *Conversations on ethics*. Oxford: Oxford University Press.