Can There Be a Preference-Based Utilitarianism?

John Broome

9.1 Introduction

John Harsanyi has made several fundamental contributions to utilitarian thinking; they are so well known that I do not need to set them out here. It was natural for him, as an economist, to present his utilitarian arguments in terms of preferences. His great influence has been a major factor in diverting the mainstream of utilitarian thinking toward a preference-based – I shall call it *preferencist* – version of utilitarianism. Preferencism is the view that good – what is good for a person and what is good overall – is determined entirely by people's preferences. However, Harsanyi himself brings into his arguments elements that are not preferencist, and I think that was inevitable. Preferences may partly determine good, but other things must enter too.

To an extent, this is obvious. If good is determined by preferences, we have to ask what determines *how* it is determined by preferences. If good is a function of preferences, what determines the functional form? Perhaps the functional form might itself be determined by preferences, but then what determines the way that happens? At some level, something other than preferences must come into the determination. In this chapter, I shall investigate what extra besides preferences is required to produce a coherent version of utilitarianism. How preferencist can utilitarianism be? It does no great harm to preferencism if nonpreferencist considerations of some sort have to be brought in from elsewhere. But it would be seriously damaging if we had to import substantive claims that make good depend on something other than preferences. Claims like these would actually conflict with preferencism.

Many of us believe preferencism is false anyway. It is often argued that we have other moral aims besides satisfying preferences. Perhaps, indeed,

---

satisfying preferences should not in itself be a moral aim at all. I am sure many of these arguments are sound, but I shall not use them in this chapter. They are unlikely to convince a preferencist utilitarian, because utilitarians, in general, and preferencist utilitarians in particular, are usually reformist. If we have other moral aims besides satisfying preferences, they may well think we should change our moral aims. For the same reason, I shall not rely on our intuitive grasp of what is good, or of what is good for a person. In any case, I doubt we have an intuitive grasp that is adequate for the purposes of utilitarianism. Utilitarianism requires good to be quantitative in a particular sense I shall specify more exactly later. It is not enough for utilitarianism that things should be ordered by their goodness, so we have concepts of better and worse. We also need a concept of how much better one thing is than another. I doubt we have a clear intuitive concept of good that is quantitative in this sense. This is something that may be up for definition; a preferencist utilitarian might plausibly claim to be defining a quantitative concept of good. So instead of relying on intuition, I am going to argue on formal grounds. This will be an internal investigation of preferencist utilitarianism, testing its internal coherence. It will be asking whether preferencist utilitarianism is possible, not merely whether it is true.

Whatever the results, they will not put the value of Harsanyi's work in doubt. Harsanyi's formal arguments are very original and very important. But I think they should be cut free from their preferencist assumptions. They are more successful when reinterpreted in nonpreferencist terms. Most of my book *Weighing Goods* is an attempt to give them a more secure interpretation. That is a sign of the value I attach to them. I think we should let the preferencism go, and keep the formal arguments.

9.2 Uncertainty

Utilitarianism contains a theory of good and a theory of right. It is characteristic of the utilitarian theory of right that rightness is derived from goodness; how one should act is determined entirely by the goodness of things. The theory of good tells us how good things are, and the theory of right tells us how to act on the basis of how good things are. This chapter is about good and not right. But I need to say a little about the utilitarian theory of right by way of introduction.

For simplicity, I shall mention only the act-utilitarian version. The simplest act utilitarianism says that, when choosing between acts, you should

choose the one that will produce the best results.[1] However, this principle is in practice useless in our uncertain world. We never know certainly what results will be produced by any of our acts. So, at the time we have to act, we can never know which act we ought to do according to this principle. In order to know how to act, we need a practical way of dealing with uncertainty.

Uncertainty can be handled within either a theory of right or a theory of good. Within the theory of right, utilitarians sometimes offer this principle: when choosing between acts, one should choose the one that gives the greatest expectation of good.[2] Daniel Bernoulli appears to have assumed this,[3] and it is a version of what I call *Bernoulli's hypothesis*. It is implausible, at least on the face of it, because it implies one should be neutral about risk to good. The act that produces the greatest expectation of good may be more risky than other options: The variance in the amount of good it leads to may be higher than for other options. If so, perhaps one should choose a safer act that gives a lower expectation of good. We should not take Bernoulli's hypothesis for granted, then. But once we give it up, it is not easy to produce a sufficiently general principle within the theory of right to handle uncertainty convincingly.

For that reason, I think uncertainty is better handled within the theory of good.[4] As a principle of right, I think utilitarians should say that, when choosing between acts, one should choose the one that will lead to the best *prospect*. Then, within their theory of good, they should have an account of the goodness of prospects. A prospect is a portfolio of possible *outcomes*, each of which might come about. The goodness of a prospect will depend on the goodness of its possible outcomes. Bernoulli's hypothesis implies specifically that it is the expected goodness of its possible outcomes. But there is room within the theory of good for a more general account of the goodness of prospects.

I wish to define outcomes in a way that excludes all uncertainty; uncertainty belongs to prospects only. This means that outcomes will have to be complete histories for the world. The description of a history will be an infinitely long conjunction. In practice, then, we shall never know what the outcome of an act has been till history has come to an end. I shall call outcomes *histories*, as a reminder of what they are. We can think of a history as a degenerate prospect: the prospect in which this history certainly occurs.

---

[1] This is G. E. Moore's version. See particularly his *Ethics*, pp. 99–101.
[2] See, for instance, Derek Parfit, *Reasons and Persons*, p. 25.
[3] See his "Specimen theoriae novae de mensura sortis."
[4] This argument is more fully spelt out in my *Weighing Goods*, section 6.1.

9.3 Additivity

To keep things simple, I am going to ignore problems that involve changes in the world's population. Given an unchanging population, one central feature of the utilitarian theory of good is that good is added across people. Utilitarians are committed to at least this:

Additive Principle for Histories: One history is better than another if and only if the total of people's good is greater in the first than in the second.

Since utilitarians need to determine when one *prospect* is better than another, they will certainly need more than this. But in this chapter I shall not need to call on any stronger additive principle.

The additive principle is about aggregating together the good of different people. Next, utilitarianism needs a theory of what determines the good of the people individually. Preferencism is such a theory; I shall come to it soon. But first I must mention an attempt to derive additivity itself from preferencism. The additive principle is part of the function through which, according to preferencist utilitarians, preferences determine overall good. Unless it can be derived from preferencism, it is a nonpreferencist element within the utilitarian story. So we need to check whether the derivation can really be done.

Harsanyi tried to derive additivity from preferencism in his article "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." His argument is founded on a mathematical proof. The conclusion of the proof is certainly a sort of additivity, though it is open to question whether it is precisely the additivity of good that is set out in the additive principle. However, I am not going to pursue this question, because there is a more definitive way to refute the argument. The premises of the proof are mutually inconsistent, so they cannot all be true. Therefore, the conclusion is unsound.

There are three premises. First the Pareto principle:

Pareto Principle for Prospects: If everyone is indifferent between two prospects, these prospects are equally good. If someone prefers one prospect to another and no one prefers the other to the one, then the one is better than the other.

This principle expresses preferencism in a pure form: It says that good depends on people's preferences, and that is all it says. Harsanyi's second premise is that the relation "better than," which appears in the Pareto principle, conforms to expected utility theory. (That is to say, this relation satisfies

the axioms of expected utility theory. Expected utility theory is normally formulated as a theory of preferences, but as a formal theory, it can be applied to other two-place relations besides preferences, including the relation of betterness.) The third is that each individual's preferences also conform to expected utility theory.

These premises are mutually inconsistent because of an empirical fact that Harsanyi ignores: People do not all agree about the probabilities of every event. Some events, such as a coin's falling heads on a particular occasion, may have objective probabilities. Harsanyi's proof of his theorem assumes that all events are like that, and furthermore that everyone knows what their objective probabilities are. This was implicit when he adopted von Neumann and Morgenstern's version of expected utility theory[5] in proving the theorem; this version assumes all probabilities are objective. But in real life, many events have no objective probability; for instance, there is no objective probability that Scotland will leave the United Kingdom. Even rational, well-informed people may assign different probabilities to events like these. It turns out that when people disagree about probabilities, Harsanyi's three premises cannot all be true.[6]

At least one of them has to go, therefore. Which should it be? Perhaps more than one. But for reasons I shall not go into here,[7] the Pareto principle definitely has to be abandoned. This is not in itself much of a blow to preferencism, because this weaker version of the Pareto principle is not compromised by the objection I have given:

Pareto Principle for Histories: If everyone is indifferent between two histories, these histories are equally good. If someone prefers one history to another and no one prefers the other to the one, then the one is better than the other.

This forms the basis of the so-called "ex post" school of welfare economics,[8] and it is a solidly preferencist principle.

A bigger loss to preferencism is that the additive principle will have to come from elsewhere. It cannot itself be derived from preferencism as Harsanyi hoped. If a preferencist is to be utilitarian, then the aggregative principle of utilitarianism will have to come from some other source besides preferencism. This need not be a deep blow to preferencism, for two reasons. First,

---

[5] See von Neumann and Morgenstern, *Theory of Games and Economic Behavior*, chapter 1.
[6] This fact has been formally proved many times. See, most recently, Philippe Mongin, "Consistent Bayesian Aggregation."
[7] See *Weighing Goods*, chapter 7.
[8] See the discussion in Hild, Jeffrey, and Risse, "Preference Aggregation after Harsanyi."

additivity may be derivable by Harsanyi's own methods, if they are suitably reinterpreted. My *Weighing Goods* develops this idea.[9] The reinterpretation could preserve important elements of preferencism, such as the Pareto principle for histories. Second, preferencism could anyhow live happily with an independently derived additive principle. The additive principle is about aggregating the good of different people, whereas preferencism is most fundamentally about the good of individual people. So the two may be coexist independently.

9.4 Preferencism as an Account of Individual Good

From now on, therefore, I shall concentrate on preferencism as an account of individual good. It is one of several competing accounts that exist within the body of utilitarian thinking. It says:

Preferencist Biconditional: One history is better for a person than another if and only if the person prefers the one to the other.

Preferencism also says that the determination in this biconditional goes from right to left. The biconditional could be true in an entirely unpreferencist way. A person's good could be determined in some way independently of her preferences, and then the person could form her preferences by always preferring histories that are better for her to histories that are worse. In that case, the biconditional would be true, but preferences would be determined by good. If a person's good is to be determined by her preferences, as preferencism requires, her preferences must themselves be independent of her good.

For one thing, this means we have to be careful about the concept of preference we adopt. One concept is the dispositional one: To prefer *A* to *B* is to be disposed to choose *A* rather than *B* when you have a choice between them. This is consistent with preferencism. But the existence of another concept is revealed by this fact: I prefer to get up early rather than waste time lying in bed on Saturday mornings, but sometimes I fail to do so. Evidently, I am sometimes not disposed to get up early, but nevertheless I prefer it. I do not prefer it in the dispositional sense, but in some other sense. In fact, I prefer it in the sense that I think it would be better for me. Thinking better is one concept of preference, but it does not suit a preferencist, because a preferencist needs preference to be independent of good. The preferencist must stick to preference as a disposition.

[9] See chapter 10 particularly.

9.5 Ideal Preferencism

The version of preferencism expressed in the preferencist biconditional is too pure for almost everyone. People's preferences are often hasty, badly thought out, ill informed, inconsistent, and in various other ways defective. Even hard-line preferencists find it implausible that a person's good should be determined by such defective preferences. Most preferencists rely on preferences that are idealized in one way or another: well informed, settled in a cool hour, made mutually consistent, and so on. This gives us:

One history is better for a person than another if and only if in ideal conditions the person would prefer the one to the other.

The notion of "ideal conditions" then needs to be spelled out. However, this improved claim also seems implausible, even before spelling it out. What a person would prefer in ideal conditions might perhaps be good for her in those conditions. But what would be good for her in those conditions might be different from what is good for her in her actual unideal conditions. If you were in a cool hour, a quiet cup of coffee might be good for you, whereas as things are you need a stiff drink. To fix this problem, we have to imagine the person, in her ideal conditions, forming preferences on behalf of herself in her actual unideal conditions. We get:

Ideal Preferencist Biconditional: One history is better for a person than another if and only if the person would in ideal conditions prefer the one to the other on behalf of herself as she is.

Let us stick with this form of the biconditional. By good fortune, it cuts through another difficulty that afflicts the original preferencist biconditional. People often have altruistic preferences: They are disposed to make choices on behalf of someone else rather than themselves. These preferences evidently do not determine what is good for themselves. But now we are picking out only the preferences they have on behalf of themselves, so we are ignoring altruistic preferences.

    Once again, the determination has to go from right to left. This requirement is now not so easy to secure.[10] Ideal conditions are likely to include the condition that the person thinks about her preference. But preferencists cannot allow her to think about it in a way that presumes a notion of her good. She must not ask herself which histories would be better for her than

[10] This objection is developed by James Griffin, *Well-Being*, p. 17.

which, and determine her preferences on that basis. Instead her thinking must presumably proceed something like this. She must represent the alternative histories to herself as accurately as she can, and then just allow herself to end up preferring one or the other. This is not the most plausible model of thinking, but it is the one the preferencist must rely on.

For brevity, from now on the only preferences I shall mention are those a person would have in ideal conditions on behalf of herself as she is. I shall call these *ideal preferences*. Even when I simply say "preference," it is to be understood this way.

9.6 A Quantitative Concept of Good

The preferencist biconditional is not enough for utilitarian purposes. For each person, it determines what is better for her than what; it orders things according to their goodness for her. But a utilitarian needs more than an order; she needs a quantitative concept of good. Otherwise, the additive principle could not be applied; we could not make sense of the *total* of people's good. We must have a concept of quantities, or *degrees*, of good for a person. To cut a long story short, these degrees must be co-cardinal. This means that ratios of differences of good must be determinate both for a single person and between people. (In general, it is not enough for differences of good simply to be ordered as greater or less; their ratios must be determinate.) How can this be achieved on a preferencist basis?

Evidently, we must have a concept of degree or strength or intensity of preference that is also measured on a co-cardinal scale. That is to say, first, the degree to which a person prefers one history to another must be comparable to the degree to which she prefers a third history to a fourth. Second, this degree must also be comparable to the degree to which another person prefers one history to another. The comparability must be ratio-comparability, which means we can attach meaning to statements like "this preference is twice as strong as this one." How can this much comparability of degrees of preference be achieved?

It cannot be taken for granted. Many authors treat preferencism as the view that one should maximize the amount of satisfaction of people's preferences.[11] But this takes for granted a quantitative notion of preferences, which we are not entitled to without work. If preferencism is to progress beyond the preferencist biconditional, work has to be done.

---

[11] For instance, Brian Barry, "Rationality and Want-Satisfaction."

The question is conceptual. We must ask, What *concept* of degree of preference do we have, or can we construct, that satisfies the requirements. Having done that, we may then be up against the epistemological question of how we can find out what the degree of a particular preference is. The epistemological question may turn out easy or difficult, depending on what the appropriate concept of degree of preference turns out to be. But in any case, it is not the question I have to answer now. I am concerned with the ethical question of what makes histories good or bad. Given an answer to that, there will then be the subsidiary epistemological question of how we find out which histories are good or bad. I am not concerned with that.

Many authors have assumed that the only question is the epistemological one. R. M. Hare makes this assumption explicit.[12] He does not deal with the conceptual question, because he takes a particular concept of degree of preference for granted. He does not say explicitly what it is, but it is revealed by his argument. He says, "What I am going to discuss is the interpersonal comparison of degrees or strengths of preferences," but immediately beforehand he has said that the problem concerns "our knowledge of other people's experiences."[13] Evidently, then, he takes a degree of preference to be an experience. But degrees of preference conceived as experiences are plainly inadequate for our purposes. I dare say we have experiences associated with the degrees of some of our preferences; occasionally I experience strong longings and more occasionally weaker ones. But a huge multitude of preferences is needed to construct a measure of my good, and most of them give me no experiences whatsoever. I have a preference for being paid £120,000 annually rather than £119,950, and a preference for being paid £119,950 rather than £119,900, but I do not have time to experience these preferences. Just because we have so many preferences, most of them must be what Hume called "calm passions," "which, tho' they be real passions, produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation."[14] Hare sometimes confuses the degree of a preference between one option and another with the difference in the experiences that will result if one or the other option comes about. But that is to abandon preferencism for hedonism. Hare's work illustrates how important it is to get clear about our quantitative concept of preference before coming to epistemology.

[12] Hare, Moral Thinking, p. 117.
[13] Both quotations from *Moral Thinking*, p. 117.
[14] David Hume, *A Treatise of Human Nature*, book 2, part 3, section 3.

So, the question is conceptual: What ratio-comparable concept of degree of preference do we have or can we construct? Once we have a suitable concept, the preferencist utilitarian can use it to give us a ratio-comparable concept of degree of goodness. She can adopt the following principle, which is complicated to formulate but obviously what she requires:

Preferencist Principle: Let A, B, C, and D be histories. Let $g_i(A)$ and $g_i(B)$ be the goodnesses of A and B for a person i. Let $g_j(C)$ and $g_j(D)$ be the goodnesses of C and D for a person j. Then the ratio $\{g_i(A) - g_i(B)\}/\{g_j(C) - g_j(D)\}$ is equal to the ratio of i's degree of preference for A over B to j's degree of preference for C over D.

To establish degrees of good for a single person, we need only this extract from the preferencist principle:

Intrapersonal Preferencist Principle: Let A, B, C, and D be histories. Let $g(A)$, $g(B)$, $g(C)$, and $g(D)$ be their respective goodnesses for a person. Then the ratio $\{g(A) - g(B)\}/\{g(C) - g(D)\}$ is equal to the ratio of the person's degree of preference for A over B to her degree of preference for C over D.

9.7 The Expectational Concept

I have ruled out Hare's experience concept of degrees of preference. Expected utility theory supplies a better candidate concept, which provides comparability for a single person. Expected utility theory suggests that the degrees of a person's preferences about histories can be given by the person's preferences about uncertain prospects. The idea is this. Suppose the person prefers history A to B and history B to C. But suppose she is indifferent between B for sure and a gamble giving her either A or C at odds of one to two (that is to say, a gamble giving a 1/3 probability to A and a 2/3 probability to C). In effect, she is willing to accept one chance of making a gain from B to A in exchange for two chances of making a loss from B to C. Since she is willing to accept this gamble, the suggestion is that we should take her degree of preference for A over B to be twice her degree of preference for B over C. Developing this idea generally, expected utility theory supplies a way of constructing a complete scale of degrees of preference. It assigns a value called a *utility* to each history. The difference between the utility of one history and the utility of another is the degree to which the first is preferred to the second.

This certainly supplies a workable concept of degree of preference. I shall call it the *expectational* concept. There are alternatives. Any increasing transform – the square, for instance – of utilities measured this way provides a rival concept of degree. But there is something to be said for the expectational concept as opposed to these others. The use of probabilities provides a natural analogue of a pair of scales for measuring the strength – analogous to weight – of preferences. In the example, two chances of the loss from *B* to *C* balance the scales against one chance of the gain from *B* to *A*, so we naturally take the preference for the gain to be twice as strong as the preference against the loss. The rival concepts are less natural. Compare our concept of physical weight. Any increasing transform of weight could supply a rival concept of weight, but it would be less natural than our present concept. We use our concept because it has the natural and convenient feature that two objects each weighing one pound balance in a scale against one object weighing two pounds.

The expectational concept of degree is the most natural, but it is not forced on us by preferences alone. Preferences by themselves do not determine a concept of degree. The expectational concept is derived from preferences together with an idea of naturalness. So in adopting it, we are once more adding something to preferencism. How significant is this addition? That depends on the effect it has on our idea of good. If we adopt this concept of degrees of preference, the intrapersonal preferencist principle draws from it a corresponding concept of degrees of goodness for a person. I shall call it the expectational concept of good. Is it acceptable? Several authors have objected that it is not, or at least not necessarily. Indeed, this might be called the standard objection to Harsanyi's argument.[15]

I explained that many other concepts of degrees of preference are available. Each can pass over into an alternative concept of degrees of goodness. According to the standard objection, there is no reason to prefer one concept to another. This objection can be reinforced with another. If we adopt the expectational concept of good, it follows that, when faced with a choice between prospects, the person always (in ideal conditions) prefers the one with the greatest expectation of her good. This is Bernoulli's hypothesis again, in a different form. I have already said that Bernoulli's hypothesis is not very plausible on the face of it, because it implies risk neutrality about good. So this is a further objection.

---

[15] See, for instance, John Roemer, "Harsanyi's Impartial Observer Is *Not* a Utilitarian", Amartya Sen, "Welfare Inequalities and Rawlsian Axiomatics"; and John Weymark, "A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism".

I am not convinced by the standard objection. We have a reason to prefer the expectational concept of degree of preference to others: It is more natural. This reason carries over to expectational degrees of goodness. The preferencist may reasonably say she is constructing a quantitative concept of good for a person, and this is the one she is going to construct. If we had a clear prior concept of degrees of good for a person, which was different from the expectational one, we could use it against the expectational concept. But the preferencist may say we do not. I agree with her about that. I believe our concept of degrees of good is not immediately intuitive, and needs to be constructed in some way. If we are to go any way with the preferencist, I do not think we can deny her this construction.

The objection to Bernoulli's hypothesis also rests on a presumed prior quantitative concept of good. Since I doubt we have one, I doubt the objection succeeds. A preferencist may plausibly say that Bernoulli's hypothesis is true because our quantitative concept of a person's good is constructed in such a way that the person is risk neutral about it.

Adopting the preferencist's concept is not merely a technical matter. It has concrete consequences within utilitarianism, because it helps to determine how we ought to act: We ought to maximize the total of people's good conceived this way. So the preferencist's idea of naturalness has moral consequences. This is exactly what she intends. She believes that people's preferences, together with the most natural concept of degrees of preference, determine how we should act. If we had an alternative intuitive concept, which gave us an alternative intuition about how we should act, we could use it against her. But we do not.

9.8 Interpersonal Comparability

In sum, I think the preferencist utilitarian can survive the standard objection. Her real problem is over comparisons between people. Can she produce a concept of degree of preference that is comparable between people to the required extent?

I shall assume from now on that we have already adopted the expectational concept of degrees of preference for each person. This is cardinal; it has all the intrapersonal comparability that is required. Only one thing more is required to give full co-cardinality: Each person's degree must be made ratio-comparable with each other person's. In effect, we have to pick a unit of degree of preference for each person. Since degrees of preference are measured by utility differences, we have to make utility differences comparable between people.

The leading contender for a preferencist way of making degrees comparable is the idea of *extended preferences*. We are assuming people have preferences between histories. For instance, I prefer a history where I teach philosophy to one where I teach economics. People may also have preferences between alternatives of the form: having the characteristics of a particular person and living in a particular history. For instance, I prefer having my characteristics and living in a history where I teach philosophy to having the characteristics of an economist and living in a history where I teach economics. Harsanyi calls preferences like these *extended preferences*. He calls the objects of these preferences *extended alternatives*. Each is a pair: a set of personal characteristics together with a history. I shall call it a *life*.

Suppose I have preferences over all extended alternatives. Then my preferences will rank all the possible lives of each person; they will compare the lives of different people. Suppose furthermore that I have preferences over uncertain prospects made up of extended alternatives. Then these will determine degrees of preference in the way I have described. Because everyone's life is included within my preferences, these degrees will be comparable between different people's lives. Here are interpersonally comparable degrees of preference, in a sense.

However, they are *my* preferences only: my preferences between different sorts of lives. Other people will have different extended preferences. Because of this, extended preferences cannot give us an interpersonal scale on grounds of preference alone. We would have to choose some particular person's preferences to go on, and that could scarcely be done on a preferencist basis. Indeed, presumably it could not be done on any good basis at all. But Harsanyi and others think they have a way of overcoming this problem.[16] They claim that, once we understand the idea of extended preferences properly, we shall see that everyone has the same extended preferences as everyone else. Extended preferences are universal. Consequently, there is a firm preferencist basis for making interpersonal comparisons of degrees of preference.

I am sure this is wrong. There is no reason why people should all have the same extended preferences, and many reasons why they should not. One reason why not is that people have different values. Their values will help to determine their preferences between different lives, so these preferences will differ. For instance, I value philosophy more highly than economics,

---

[16] This argument appears most clearly set out in John Harsanyi, *Rational Behavior and Bargaining Equilibrium*, pp. 58–59. See also Kenneth Arrow, "Extended Sympathy and the Possibility of Social Choice."

so I prefer working as a philosopher and having the characteristics of a philosopher to working as an economist and having the characteristics of an economist. I imagine many economists might have the opposite preference.

To be sure, when comparing my life with an economist's, I must do it properly. In deciding whether I prefer the life and characteristics of an economist, I am supposed to take account of everything that goes with them, including having the values of an economist. I must recognize that if I had the characteristics of an economist, I would value the life of an economist. But it is my extended preferences we are talking about, not the economist's extended preferences. As it happens, I prefer not to have the values of an economist. That is one reason I prefer not to be an economist.

We should also make sure we are dealing with ideal preferences. But when two people's extended preferences disagree, neither's need be less than ideal. Each person's preferences may be fully considered and so on. At least a preferencist must think that. For a preferencist, the standard of idealness for ideal preferences cannot be so stringent as to demand that different people's values coincide. Is there a true answer to the question of whether an economist's life is better or worse than a philosopher's? Suppose there is not. In that case, even if we were in such ideal conditions that we knew everything that is true, our values need not coincide. Alternatively, suppose there is a true answer. Then perhaps in ideal conditions our preferences would coincide because they would conform to the truth. But in that case preferencism would be false. Our ideal preferences would be determined by the truth of which life was better, whereas preferencism requires the determination to be the other way round.

So it seems the extended preferences of different people need not coincide. Yet, Harsanyi offers two arguments intended to show they must coincide. One is explicit; the other implicit. The explicit argument starts off from the correct observation that if people have different extended preferences, there is a causal explanation of why they do. Of course there must be some causal explanation of why I value philosophy, and why economists value economics (if they do). Suppose it is the star signs we were born under. Harsanyi claims that if we include this cause among the objects of our preferences, then our preferences will all be the same. But this is false. Perhaps we care about star signs and perhaps we do not, but at any rate, there is no reason why our preferences about them, or about anything else, should coincide. Harsanyi was led to his conclusion by a technical mistake, which I have said enough about elsewhere.[17]

---

[17] See my "A Cause of Preference Is Not an Object of Preference" and "Extended Preferences."

It is true that if we were all in the same causal situation, we would all have the same preferences. But we are not. Perhaps we could pick out some privileged causal situation, and base our interpersonal comparisons on the extended preferences we would have in that situation. Harsanyi sometimes seems to have in mind for this role a sort of causally empty situation, where we have been acted on by no causes apart from our bare human nature. He suggests we should use the preferences we would have in this causally empty situation. This is his second, implicit, argument for the claim that extended preferences are universal. But it is surely a fantasy to suppose we could have preferences determined by bare human nature.[18]

When he comes to a concrete case, Harsanyi has a quite different way of proceeding. He says,

> For example, if I want to compare the utility that I would derive from a new car with the utility that a friend would derive from a new sailboat, then I must ask myself what utility I would derive from a sailboat if I had taken up sailing for a regular hobby as my friend has done, and if I could suddenly acquire my friend's expert sailing skill, and so forth.[19]

Harsanyi evidently proposes to estimate how well off he would be if he had acquired a new sailboat and all his friend's sailing skills. He seems to be planning to form his extended preferences on the basis of an estimate of the benefits of leading a life like his friend's. This implies that the benefits of this life are determined in advance of Harsanyi's preferences. It is an anti-preferencist view. It presupposes an idea of people's good that is independent of preferences. This is why I said in Section 9.1 that Harsanyi's own theory contains nonpreferencist elements.

9.9 Evolutionary Equilibrium

Ken Binmore offers a new theory developing the idea of extended preferences.[20] He argues that causal processes of social evolution determine our extended preferences. In the long run,[21] he argues, extended preferences will converge. This provides a potential new basis for preferencism. The difficulty with using extended preferences to provide interpersonal comparisons is that people's extended preferences differ. But Binmore supplies an argument to say they will not differ in the long run. No doubt we shall always

---

[18] For a discussion, see M. Kaneko, "On Interpersonal Utility Comparisons."
[19] Harsanyi, *Rational Behavior and Bargaining Equilibrium*, p. 59.
[20] Binmore, "Naturalizing Harsanyi and Rawls."
[21] Technically this is Binmore's *medium term*.

find some disagreements in our actual extended preferences: I suggested mine differ from an economist's. But Binmore would think these are minor deviations that exist only because social evolution has not had time to iron them out. From a broad viewpoint, adopting a long timescale, extended preferences converge.

Let me describe Binmore's view in a little more detail. His argument is set in a special strategic situation, where people regularly negotiate with each other behind an imagined veil of ignorance. People negotiate in pairs, to distribute goods between each other. Behind the imagined veil, neither person is supposed to know whose position she will occupy once the negotiation is completed; it might be her own position with her own characteristics or the other person's position with the other person's characteristics. In these conditions, the two settle on a distribution on the basis of their extended preferences. These preferences are formed by social evolution. This means that people tend to copy the attitudes of people they see doing well in their negotiations. Binmore argues that this process will drive us all to the same extended preferences in the long run. To be more precise, we will all make the same interpersonal comparisons of degrees of preference.

A more specific outcome of Binmore's argument is surprising. It turns out that in the long run we will assign high degrees of preference to people who have a lot of bargaining power. Let us suppose everyone prefers a day of sunshine to a day of rain. We will assign a higher degree to this preference when it belongs to a powerful person than we do when it belongs to a less powerful person. We shall suppose powerful people have more intense preferences. If we feed this conclusion into the preferencist principle, we shall conclude that powerful people tend to get more benefit from good things than less powerful people do. Consequently, if goods are distributed on a utilitarian basis, the lion's share will go to the powerful. Naturally, these same people will also get the lion's share if the goods are distributed by a free-for-all. In the long run, utilitarianism will reproduce what would have been the result of a free-for-all. Binmore derives this conclusion by mathematics, but he does not offer an intuitive explanation of why it happens.

What does all this do for preferencism? At first, it seems to give it support. Preferencism was laboring under the difficulty that preferences did not seem to provide a basis for interpersonal comparisons of degrees of preference. Extended preferences were supposed to do the job, but different people have different extended preferences, and there are no preferencist grounds for choosing between them. Now Binmore suggests these differences are unimportant. They are temporary only. Our extended preferences will converge in the long run because evolutionary processes will make sure they do. So we

can perhaps ignore the differences. Moreover, the preferences we are converging on are determined entirely by blind causal forces. They contain no taint of a nonpreferencist theory of good. All this is good for preferencism.

But actually this very ethical neutrality prevents Binmore's argument from supporting preferencism. Binmore calls his theory "naturalistic." I believe he means to say it is a natural history of ethical beliefs. He thinks people's extended preferences are a natural feature of people, determined by natural, causal processes, and he aims to give an explanation of these processes. Because they determine extended preferences, these natural processes determine degrees of preference that are comparable between people. They will lead people to make interpersonal comparisons of degrees of good, corresponding to the degrees of their preferences. That is to say, these evolutionary processes will cause people to have certain beliefs about how good or bad things are for people. In the end, they will also lead people to have particular beliefs about how they and others ought to act. Let us suppose they will lead them to utilitarian beliefs, with interpersonal comparisons determined in the way described by Binmore.

A successful natural history of ethics might explain why people will believe they ought to maximize the total of people's good. It may also show that their notion of good will derive from preferences. But a preferencist utilitarian needs something quite different. She needs a demonstration that people ought to maximize the total of people's good, where people's good is determined by their preferences. A natural history of ethics gives no support to these claims whatsoever. It may tell us what people will believe, but it does so in a way that gives no grounds for their beliefs. This type of naturalism passes ethics by. It is irrelevant to preferencism, since preferencism is an ethical theory.

9.10 Conclusion

I conclude that preferencist utilitarianism fails. Preferencism cannot generate a concept of good solid enough to make sense of interpersonal comparisons of good. Interpersonal comparisons can only be achieved by means of a different, nonpreferencist theory of good.

References

Arrow, K. J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review, Papers and Proceedings* 67, 219–225; reprinted in his *Collected Papers,* Vol. 1: *Social Choice and Justice*, Harvard University Press, Cambridge, MA, 1983, pp. 147–161.

Barry, B. 2008. Rationality and want-satisfaction. *In Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 281–299.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, vol. 5; trans. Louise Sommer, Exposition of a new theory on the measurement of risk. *Econometrica* 22 (1954), 23–36.

Binmore, K. 2008. Naturalizing Harsanyi and Rawls. *In Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 303–333.

Broome, J. 1993. A cause of preference is not an object of preference. *Social Choice and Welfare* 10 (1993), 57–68.

Broome, J. 1998. Extended preferences. In *Preferences*, ed. C. Fehige and U. Wessels, de Gruyter, Berlin, pp. 279–296.

Broome, J. 1991. *Weighing Goods*. Blackwell, Oxford.

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

Hare, R. M. 1982. *Moral Thinking: Its Levels, Method and Point*. Oxford University Press, Oxford.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–321; reprinted in his *Essays on Ethics, Social Behavior, and Scientific Explanation*, D. Reidel, Dordrecht, 1976, pp. 6–23.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge.

Hild, M., Jeffrey, R., and Risse, M. 2008. Preference aggregation after Harsanyi. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 198–217.

Hume, D. 1978. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch. Oxford University Press, Oxford.

Kaneko, M. 1984. On interpersonal utility comparisons. *Social Choice and Welfare* 1, 165–175.

Mongin, P. 1995. Consistent Bayesian aggregation. *Journal of Economic Theory* 66, 313–351.

Moore, G. E. 1966. *Ethics*, 2nd ed. Oxford University Press, Oxford.

Parfit, D. 1984. *Reasons and Persons*. Oxford University Press, Oxford.

Roemer, J. Harsanyi's impartial observer is *not* a utilitarian. In *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. M. Fleurbaey, M. Salles, and J. A. Weymark, Cambridge University Press, Cambridge, pp. 129–135.

Sen, A. 1976. Welfare inequalities and Rawlsian axiomatics. *Theory and Decision* 7, 243–262.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd ed. Princeton University Press, Princeton, NJ.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge University Press, Cambridge, 255–320.