Interpersonal Comparison of Utility

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT, UK

Interpersonal Comparison of Utility

by Ken Binmore[1]

> 'Tis vain to talk of adding quantities which after the addition will continue to be as distinct as they were before; one man's happiness will never be another man's happiness: a gain to one man is no gain to another: you might as well pretend to add 20 apples to 20 pears.
>
> Jeremy Bentham

1 Introduction

There are at least as many views on how the welfare of individuals should be compared as there are authors who write on the subject. An indication of the bewildering range of issues considered relevant in the literature is provided by the book *Interpersonal Comparisons of Well-Being* (Elster and Roemer [8]). However, I plan to interpret the topic narrowly. Although I shall review some traditional approaches along the way, my focus will be on the questions:

What do modern economists mean when they talk about units of utility?

How can such utils be compared?

It is widely thought that that the answer to the second question is that utils assigned to different individuals cannot sensibly be compared at all. If this were true, I share the view expressed by Hammond [11], Harsanyi [14] and many others that rational ethics would then become a subject with little or no substantive content.

2 What is Utility?

This section offers a brief sketch of the history of utility theory with a view to explaining the origins and tenets of the modern theory.

The word *utility* has always been difficult. Even the arch-utilitarian Jeremy Bentham [2] opens his *Principles of Morals and Legislation* by remarking that his earlier work would have been better understood if he had used *happiness* or *felicity* instead. The emergence of modern utility theory has only served to multiply the philosophical confusion. For example, Amartya Sen [30] denies that John Harsanyi

[1] Some of this material is a reworking of sections of my book *Natural Justice* (Binmore [5] and its earlier two-volume incarnation *Game Theory and the Social Contract* (Binmore [3, 4]). My *Playing for Real,* which is forthcoming with Oxford University Press, contains many details omitted in this article.

[15] can properly be called a utilitarian, because he interprets utility in the modern sense of Von Neumann and Morgenstern rather than in Bentham's original sense.[2]

Bentham's position, later taken up by John Stuart Mill, is that utility should be interpreted as the balance of pleasure and pain experienced by an individual human being. In the fashion of the times, he gave long lists of the phenomena that he felt relevant when estimating a person's utility from the outside.

Victorian economists took up Bentham's idea and incorporated it into their models without paying much attention to its doubtful philosophical and psychological foundations. However, once economists discovered (in the "marginalist revolution" of the early part of the twentieth century) that they did not need to attribute utility functions to economic agents in order to prove most of the propositions that seemd important at the time, all of the baggage on utility theory inherited from the Victorian era was swept away. By the late 1930s, it had become fashionable for economists to denounce cardinal utility theory as meaningless nonsense.[3] Even at this late date, Lionel Robbins [26] is still sometimes quoted to this effect. However, even at the time Robbins was making his name by denouncing classical utility theory, Von Neumann and Morgenstern [34] were creating the beginnings of the entirely sensible modern theory to which the main part of this paper is devoted.

But Von Neumann and Morgenstern's theory has not satisfied everybody, and now that those who lampooned Bentham's empirical approach are no longer with us, his ideas have been revived as the new field of "happiness studies" (Layard [21]). Amartya Sen offers lists of "capabilities" that might reasonably be expected to promote an agent's well-being. (See Cohen [6]). There are even neuroscientists who are willing to contemplate the possibility that some kind of metering device might eventually be wired into a brain to measure how much pleasure or pain a person is experiencing. Now that everybody knows of the experiments in which rats press a lever that excites an electrode implanted in a "pleasure center" in their brains to the exclusion of all other options (including food and sex), this idea is perhaps not as wild as it once would have seemed. These and other revivals of Bentham's psychological theory are not necessarily inconsistent with the orthodox approach that I refer to as modern utility theory, but it is important for philosophers to recognize that the modern theory does not depend on any of the psychological or physiological assumptions built into these modern counterparts of the theory of Bentham and Mill.

[2] I suppose it is idle to suggest that we start using the word *felicity* for Bentham's psychological notion, in order to distinguish it from the very different manner in which modern economists use the word *utility*. Philosophers sometimes speak of 'preference satisfaction' when referring to the modern usage, but I suspect they seldom understand how radical the change in attitude has been.

[3] A cardinal utility scale operates like a temperature scale, with utils replacing degrees. It is normally contrasted with an ordinal utility scale, in which the amount by which the utility of one outcome exceeds the utility of another outcome is held to be meaningless.

3 Modern Utility Theory.

Critics of modern utility theory usually imagine that economists still hold fast to the naive beliefs about the way our minds work that are implicit in the work of Bentham and his modern emulators, but orthodox economists gave up trying to be psychologists a long time ago. Far from maintaining that our brains are little machines for generating utility, the modern theory of utility makes a virtue of assuming $nothing\ whatever$ about what causes our behavior.

This doesn't mean that orthodox economists believe that our thought processes have nothing to do with our behavior. We know perfectly well that human beings are motivated by all kinds of considerations. They care about pleasure, and they care about pain. Some are greedy for money.[4] Others just want to stay out of jail. There are even saintly people who would give away the shirt off their back rather than see a baby cry. We accept that people are infinitely various, but we succeed in accommodating their infinite variety within a single theory by denying ourselves the luxury of speculating about what is going on inside their heads. Instead, we pay attention only to what we see them doing.

The modern theory of utility therefore abandons any attempt to explain why people behave as they do. Instead of an explanatory theory, we have to be content with a descriptive theory, which can do no more than say that a person will be acting inconsistently if he or she did such-and-such in the past, but now plan to do so-and-so in the future.

Such a theory is rooted in observed behavior. It is therefore called a theory of "revealed preference", because the data we use in determining what people want is not what they say they want—or what paternalists say they ought to want—but our observations of what they actually choose when given the opportunity.

When using a revealed preference theory, one must beware of the causal utility fallacy, which says that decision-makers choose a over b because the utility of a exceeds that of b. Modern utility theory does not allow such a conclusion. In the psychological theory of Bentham and Mill, one may certainly argue that a person's choice of a over b is $caused$ by the utility of a exceeding that of b. But in modern utility theory, the implication goes the other way. It is because the preference $a \succ b$ has been revealed that we choose a utility function satisfying $u(a) > u(b)$.

For people to behave $as\ though$ their aim were to maximize a utility function, it is only necessary that their choice behavior be consistent. To challenge the theory, one therefore needs to argue that people behave inconsistently, rather than that they don't really have utility generators inside their heads.

Rationality as consistency. Modern utility theory famously began when Oskar Morgenstern turned up at Von Neumann's house in Princeton one day in the early

[4] Bentham [2] suggests at one point that we should measure felicity in money, as in modern cost-benefit analysis, but who would want to argue that an extra dollar is worth the same to a billionaire as to a beggar?

forties complaining that they didn't have a proper basis for the payoffs in the book on game theory they were writing together. So Von Neumann invented a theory on the spot that measures how much a rational person wants something by the size of the risk he is willing to take to get it.

Critics sometimes complain that a person's attitude to taking risks is only relevant when gambling, and so Von Neumann's theory is irrelevant to the kind of moral situations to which utilitarians apply their ideas, but it is hard to think of a more fundamental moral question than: who should bear what risk? Utilitarians who want to use the kind of insurance argument employed by Edgeworth [7] or Harsanyi [15] in defense of their position certainly have no choice but to accept that attitudes to risk are basic to their approach. Nor is there any lack of support from traditional sources. As it says in the Book of Proverbs: it is the lot that causeth contentions to cease.

The rationality assumptions built into Von Neumann's theory simply require that people make decisions in a consistent way, but his conclusions are surprisingly strong. Anyone who chooses consistently in risky situations will look to an observer as though he or she were trying to maximize the expected value of something. This abstract "something" is what is called utility in the modern theory. To maximize its expected value is simply to take whatever action will make it largest on average.

There are philosophers who follow Kant in claiming that rationality should mean more than mere consistency, so that some utility functions can be dismissed as being less rational than others. However, modern economists follow Hume in treating reason as the "slave of the passions". There can then be nothing irrational about consistently pursuing any end whatever. As Hume extravagantly observed, he might be criticized on many grounds if he were to prefer the destruction of the entire universe to scratching his finger, but his preference could not properly be called *irrational*, because (contra Kant) rationality is about means rather than ends.

Determining a person's utility. Von Neumann's theory makes it easy to find a utility function that describes a rational person's behavior if enough data is available on the choices he or she has made in the past between risky prospects.

Pick two outcomes, $\mathcal{W}$ and $\mathcal{L}$, that are respectively better and worse than any outcome that we need to discuss. (One can think of $\mathcal{W}$ and $\mathcal{L}$ as winning or losing everything that there is to be won or lost.) These outcomes will correspond to the boiling and freezing points used to calibrate a thermometer, in that the utility scale to be constructed will assign 0 utils to $\mathcal{L}$, and 100 utils to $\mathcal{W}$.

Suppose we now want to find David Hume's Von Neumann and Morgenstern utility for scratching his finger. For this purpose, consider a bunch of (free) lottery tickets in which the prizes are either $\mathcal{W}$ or $\mathcal{L}$. As we offer Hume lottery tickets with higher and higher probabilities of getting $\mathcal{W}$ as an alternative to scratching his finger, he will eventually switch from saying no to saying yes. If the probability of $\mathcal{W}$ on the lottery ticket that makes him switch is 73%, then Von Neumann and Morgenstern's theory says that scratching his thumb should count as being worth

73 utils to Hume. Each extra percentage point added to the indifference probability therefore corresponds to one extra util.

As with measuring temperature, it will be obvious that we are free to choose the zero and the unit on the utility scale we construct however we like. We could, for example, have assigned 32 utils to $\mathcal{L}$, and 212 utils to $\mathcal{W}$. One then finds how many utils a scratched finger is worth on this new scale in the same way that one converts degrees Celsius into degrees Fahreneheit. So a scratched thumb worth 73 utils on the old scale is worth 163.4 utils on the new scale.

My guess is that Bentham would have been delighted with the mechanical nature of Von Neumann and Morgenstern's theory, which reduces evaluations of individual welfare to ticking off boxes on a simple Gradgrindian questionnaire, but he would also probably have made the same mistake as many economists in over-estimating the extent to which real people make their choices in a consistent manner. Although the utility theory of Von Neumann and Morgenstern performs at least as well as any comparable alternative proposed by the new school of behavioral economists, it is only the rather poor best of a bad lot when it comes to predicting the behavior of laboratory subjects.

Intrapersonal comparison. Why does a rich man hail a taxicab when it rains while a poor man gets wet? Economists answer this traditional question by making an intrapersonal comparison. They argue that an extra dollar in Adam's pocket would be worth more to him if he were poor than it would be if he were rich. But how are we to measure such increments in well-being? The unit of measurement certainly cannot be the dollar because, for most of us, an extra dollar becomes less valuable the more of them we have—a phenomenon that economists refer to as the decreasing marginal utility of money.

The answer offered by the Von Neumann and Morgenstern theory is that one can measure the well-being of a person with decreasing (or increasing) marginal utility of money by counting the number of utils by which his total utility is increased when he gains an extra dollar. Because a rational decision-maker in the Von Neumann and Morgenstern theory acts as though maximizing expected or average utility, he behaves as though each util is "worth" the same as any other.

Risk aversion. The Von Neumann and Morgenstern utility function of a person with decreasing marginal utility for money has a concave shape (chords drawn to the curve lie underneath the curve). It follows that such a person is risk averse—that is to say, he prefers to get a sum of money for certain than to participate in a risky enterprise that yields the same amount on average. A risk-loving person (with a convex Von Neumann and Morgenstern utility function) will prefer the risky enterprise. A risk-neutral person will always be indifferent when offered such a choice. In spite of what is commonly said about the beliefs of economists, it is only in the risk-neutral case that they believe in identifying a util with a dollar.

There is sometimes discussion in the philosophical literature about the rationality

of "prudent behavior", but the modern theory of utility denies that the extent to which a person is risk averse is a rationality question. As with David Hume's finger, a person's attitudes to taking risks are deemed to be part of his preferences. Epictitus said that one should not say of a man who drinks much wine that he drinks too much, but only that he drinks much wine. We similarly say of a man who risks a lot in gambling dens, not that it is irrational for him to risk so much, but that he behaves as though he has risk-loving preferences.

Gambling and gut feelings. In discussing Rawls' [24] arguments for the use of the maximin criterion in the original position, Kukathas and Pettit [20, p.40] dismiss the maximization of expected Von Neumann and Morgenstern utility as "the gambling strategy". However, numerous economics textbooks to the contrary, a Von Neumann and Morgenstern utility function does not measure how much a person likes or dislikes gambling as an activity. Indeed, as Harsanyi [16] points out, Von Neumann and Morgenstern's assumptions *do not apply* if a person actually derives pleasure or distress from the process of gambling itself.[5]

Nor is it true that the maximin criterion employed by Rawls [24] in evaluating the decision problem faced by an individual in the original position corresponds to the case of infinite risk aversion. The most concave utility function for money would assign zero utils to no dollars and one util to any positive number of dollars—corresponding to the case in which someone cares only about not ending up with nothing at all. But Rawls certainly did not intend that such utility functions should be assigned to the citizens of his ideal society.

Such misunderstandings of the concept of risk aversion derive in part from the entrenched use of the language of the casino and the racetrack in discussing the Von Neumann and Morgenstern theory. However, the rational folk to whom the theory genuinely applies will see no point in taking vacations in Las Vegas. A better exemplar of the theory is a Presbyterian minister considering how to value his house or car. He will not regard the possibility that his house might burn down or his car be stolen as a possible source of excitement. He will make a sober assessment of the probabilities and of the value he places on his property, and then take out insurance to cover himself against the objectively determined risks he faces.

As an example of the power of the theory, imagine that some bullets are loaded into a revolver with six chambers. The cylinder is then spun and the gun pointed at your head. Would you now be prepared to pay more to get one bullet removed when only one bullet was loaded or when four bullets were loaded? Usually, people say they would pay more in the first case because they would then be buying their lives for certain. But the Von Neumann and Morgenstern theory says that you should pay more in the second case, provided only that you prefer life to death and more

[5] Someone who likes or dislikes the gambling process for its own sake would be unlikely to accept Von Neumann and Morgenstern's assumption that two different lotteries should be regarded as equivalent when they generate the same prizes with the same probabilities. He would prefer whichever lottery squeezed the most drama and suspense from the randomizing process.

money to less.[6]

What conclusion should be drawn from such a conflict between one's gut feelings and the recommendations of Von Neumann and Morgenstern's theory? Few people want to admit that their gut feelings are irrational and should therefore be amended—which was the reaction of the statistician Savage when trapped by the economist Allais into expressing preferences that were inconsistent with his extension of the Von Neumann and Morgenstern theory. They prefer to deny that the Von Neumann and Morgenstern assumptions characterize rational behavior.

On this subject, it is instructive to consider another informal experiment with which I have teased various experts in economics and finance. Would you prefer $\$(96 \times 69)$ or $\$(87 \times 78)$? Most prefer the former. But $96 \times 69 = 6,624$ and $87 \times 78 = 6,786$. How should we react to this anomaly? Surely not by altering the laws of arithmetic to make $96 \times 69 > 87 \times 78$! So why should we contemplate altering the Von Neumann and Morgenstern assumptions after observing experiments that show they don't match the gut feelings of the man in the street? Our untutored intuitions about statistical matters are no more trustworthy than those that lead a toddler to prefer a candy jar with a big cross-section to a rival with a larger volume. Adults learn to think twice about such matters. If the matter is sufficiently important, we may calculate a little—or perhaps read the label on the packet.

It is such second thoughts about our wants and aspirations that seem to me relevant when ethical issues are at stake. A theory based on gut feelings may perhaps be appropriate in consumer theory when studying impulse buying in supermarkets. But people are likely to be in a much more reflective mood when considering social contract issues. As the old form of English marriage service used to say, this is an occasion for making decisions "advisedly and soberly"—just as a Presbyterian minister contemplates his insurance problem.

Intensity of preference. In spite of its ancient provenance, the book *Games and Decisions* by Luce and Raiifa [23] continues to be an excellent reference for one-person decision theory. The fallacies it lists in Chapter 2 remain as relevant now as in the 1950s when the book was written. The third fallacy is of particular importance in discussions of the meaning of utilitarian welfare functions. It says that if the difference in utilities between outcomes a and b exceeds the difference in utilities between outcomes c and d, one is not entitled to deduce that the decision-maker would prefer a change from b to a to a change from d to c.

[6] Suppose that you are just willing to pay $\$X$ to get one bullet removed from a gun containing one bullet and $\$Y$ to get one bullet removed from a gun containing four bullets. Let $\mathcal{L}$ mean death. Let $\mathcal{W}$ mean being alive after paying nothing, $\mathcal{C}$ mean being alive after paying $\$X$ and $\mathcal{D}$ mean being alive after paying $\$Y$. Then $u(\mathcal{C}) = \frac{1}{6}u(\mathcal{L}) + \frac{5}{6}u(\mathcal{W})$ and $\frac{1}{2}u(\mathcal{L}) + \frac{1}{2}u(\mathcal{D}) = \frac{2}{3}u(\mathcal{L}) + \frac{1}{3}u(\mathcal{W})$. Simplify by taking $l = u(\mathcal{L}) = 0$ and $w = u(\mathcal{W}) = 1$. Then $c = u(\mathcal{C}) = \frac{5}{6}$ and $d = u(\mathcal{D}) = \frac{2}{3}$. Thus $u(\mathcal{D}) \prec u(\mathcal{C})$ and so $X < Y$. (This elegant problem is attributed to Zeckhauser by Gibbard [9] and Kahneman/Tversky [19]. Kahneman and Tversky think the example misleading on the grounds that matters are confused by the question of whether money has value for you after you are dead.)

If one wants to make such a claim it is necessary to add extra assumptions to the standard Von Neumann and Morgenstern theory. In particular, one needs to offer a definition of what it means to say that one preference is held more intensely than another. Von Neumann and Morgenstern themselves [34, p.18] suggest that Adam should be deemed to hold the first preference more intensely than the second if and only if he would always be willing to swap a lottery $\mathbf{L}$ in which the prizes a and d each occur with probability $\frac{1}{2}$ for a lottery $\mathbf{M}$ in which the prizes b and c each occur with probability $\frac{1}{2}$. To see why they proposes this definition, imagine that the citizen is in possession of a lottery ticket $\mathbf{N}$ that yields the prizes b and d with equal probabilities. Would he now rather exchange b for a in the lottery or d for c? Presumably, Adam will prefer the latter swap if and only if he thinks that b is a greater improvement on a than c is on d. But to say that he prefers the first of the two proposed exchanges to the second is to say that the citizen prefers $\mathbf{M}$ to $\mathbf{L}$.

With Von Neumann and Morgenstern's definition, Luce and Raiffa's third fallacy evaporates, but it retains its sting with other definitions of intensity of preference.

4 Interpersonal Comparison of Utility

We have seen that there is a sense in which each util that a rational Adam receives is worth the same to him as all the previous utils he has received. But what of the utils acquired by a rational Eve? In the absence of a Tiresias with experience of both roles, to whom do we appeal when asked to compare the welfare of two such different people as Adam and Eve?

Do we really need to compare utilities at all? For example, one might compare Adam's welfare with Eve's by counting their daily consumption of apples. Such a procedure has the advantage of being relatively easy to operationalize, but it is vulnerable to numerous criticisms. Perhaps Eve is rich and Adam is poor, so that she can eat as many apples as she chooses but his straitened circumstances restrict him to only one apple a day. It would then be surprising if Eve did not derive a great deal less joy from one extra apple per day than Adam.[7]

Even if everybody had equal access to apples, we would still have problems. Suppose that Adam and Eve are both poor, but Adam cares only for fig leaves. Rawls [25] is only one of many who see no difficulty here, on the grounds that fig

[7] The economists' notion of *consumer surplus* faces precisely this problem—even in the special case of quasilinear utility, for which Varian [33, p.169] describes consumer surplus as being exactly the appropriate welfare measure. An *ordinal* utility function for Adam is quasilinear if it assigns utility $a + U(f)$ to a commodity bundle consisting of a apples and f fig leaves. One can then regard U as defining a *cardinal* utility scale for fig leaves. Adam will always be ready to swap one util's worth of fig leaves for one apple, and so any util on the fig-leaf scale is exactly comparable to any other—if the standard for comparison is the number of apples that Adam will trade for it. But, who is to say that apples (or dollars) are the "appropriate" standard of comparison?

leaves can be exchanged for apples in the marketplace. One can therefore assess their relative values by quoting their price in dollars. But to grant some *a priori* legitimacy to the market mechanism begs exactly the sort of question that Rawls is seeking to answer. Even when a market for apples and fig leaves can be taken for granted, the rate at which apples are exchanged for fig leaves in this market tells us little about the relative felicity that Adam and Eve derive from consuming apples and fig leaves. Markets are driven by the relative scarcity of the goods that are traded. If Eve is a pop star, she may be able to trade a nanosecond of her labor for a kidney dialysis machine. But that does not mean that the sacrifice of a nanosecond of her time is worth the same to her as a kidney machine would be to Adam if he were suffering from kidney failure.

One can, of course, invent an index of goods in which the weights attached to fig leaves and apples somehow reflect their use-value rather than their exchange-value. However, objections are not hard to find. For example, it may be that Adam likes martinis mixed only in the proportions of 10 parts of gin to each part of vermouth, whereas the hard-drinking Eve likes them only in the proportions of 1,000 parts of gin to each part of vermouth. It then makes little sense to say that Adam and Eve are equally well-off if each are assigned 10 bottles of gin and 1 bottle of vermouth. Adam will now be able to enjoy 11 bottles of martini, whereas Eve will be able to enjoy only 10.01 bottles of martini.[8]

Such examples suggest an idea that it is familiar in the economic theory of the household. Instead of measuring Adam and Eve's welfare in terms of the raw commodities they consume, why not measure their welfare in terms of the characteristic benefits that they get from their consumption? One might ask, for example, how much health, wealth and wisdom an agent derives from any given bundle of consumption goods. Rawls [25] proposes just such a list of "primary goods". The primary goods he proposes for aggregation in an index are "the powers and prerogatives of office", "the social basis of self-respect" and "income and wealth".

Economists are not fond of theories that depend on such intangibles. However, even if Rawls' primary goods could be defined in precise terms, one would still be faced with an indexing problem. How do we weigh such primary goods against each other? Rawls seems to think this is a matter on which a broad consensus can be expected.[9] However, I believe that we cannot rely on different individuals valuing some set of primary goods in a similar fashion. Even my own dean is uncharacteristically obtuse when it comes to weighing my self-respect against the

[8]Notice that, after Adam and Eve have each been assigned the same bundle of commodities, neither will see any advantage in swapping their bundles. Economists say that an allocation of commodities with this property is *envy-free*. No interpersonal comparisons of utility need be made in identifying such an envy-free allocation. However, I hope the gin and martini example will suffice to indicate why it is wrong to deduce that interpersonal comparisons are therefore irrelevant.

[9]Rawls [24, p.94] speaks of the judgment that would be made by a representative agent. He then muddies the waters by saying that only a representative of the least-advantaged social grouping need be considered. But how does one know which group is least advantaged if one has not already decided the scale that determines advantage?

prerogatives of her office! Moral philosophy would be a great deal easier if there really were primary goods about which everybody felt pretty much the same, but to make this kind of assumption is to beg a whole set of major questions. Such attempts to evade the problem of interpersonal comparison of utility are admittedly tempting, but they lead only to confusion in the long run.

Interpersonal comparisons are impossible? As Hammond [12] documents, establishing a standard for making interpersonal comparisons of utility is widely regarded as impossible or hopelessly intractable. At the time when logical positivism was fashionable, the sound and fury raised against theories of interpersonal comparison reached an almost hysterical pitch. Echoes of the debate still haunt the economics profession today, with the result that expedients like those reviewed above continue to be invented with the aim of somehow allowing interpersonal judgments to be made without utils actually being compared. Meanwhile, a parallel movement within philosophy works hard at making a technical subject out of the notion that ethical values are supposedly incommensurable by their very nature.

I think John Harsanyi's [15] theory of interpersonal comparisons of utility defuses all these concerns by providing a clear and relatively uncontroversial approach to the subject, but I want first to look at two other approaches—not because I think them satisfactory, but to make it clear that the question isn't really whether interpersonal comparison is possible, but which of all the ways it might be done works best in a given context.[10]

Counting perception thresholds. I believe that it was Edgeworth [7] who first proposed observing how far a parameter controlling the environment of a subject needs to be changed before the subject perceives that a change has taken place. If the subject expresses a preference for low parameter values over high parameter values, the number of perceptual jumps he experiences as the parameter moves from one end of its range to the other can then be used as a measure of the intensity of his preference between the two extremes. The psychologist Luce [22] has been a modern exponent of this idea. Rubinstein [27] has also explored its implications.

If workable, such a procedure would provide an objective method of comparing utils across individuals independently of the social mores of their society of origin. But would such a comparison be meaningful? Even if it were, would it be possible to persuade people to regard it as a relevant input when making fairness judgments?

Like many men, I am not only nearsighted, I am also mildly color-blind. At the Poker table, I have to be quite careful when both blue and green chips are in use. Does it therefore follow that I get less pleasure from the use of my eyesight than someone with perfect vision? My hearing is even less reliable than my eyesight. Should those with perfect pitch therefore be assumed to take a keener pleasure in music? I have only the haziest idea of how much I am worth, while others keep

[10] My *Game Theory and the Social Contract* contains several other examples. (Binmore [4])

accounts that are accurate down to the penny. Is this relevant to how much tax we each should pay?

Zero-one rule. Similar doubts afflict another proposal that gets a better press. Hausman [17] even argues that, if there is a "correct" way to compare bounded cardinal utilities, then the zero-one rule is it.

The zero-one rule applies when it is uncontroversial that a person's individual preferences are to be measured with a cardinal utility function, usually a Von Neumann and Morgenstern utility function. If Adam and Eve agree that the worst thing that can happen for both of them separately is $\mathcal{L}$ and the best is $\mathcal{W}$, then the zero-one rule calls for their utility scales to be recalibrated so that their new Von Neumann and Morgenstern utility functions v_A and v_E satisfy $v_A(\mathcal{L}) = v_E(\mathcal{L}) = 0$ and $v_A(\mathcal{W}) = v_E\mathcal{W}) = 1$. The utility functions obtained after such a recalibration can then be compared without difficulty.

Essentially the same objections to this procedure have been made by Griffin [10], Hammond [12], Rawls [24], and Sen [29]. If Eve is a jaded sophisticate who sees $\mathcal{W}$ as only marginally less dull than $\mathcal{L}$, whereas Adam is a bright-eyed youth for whom the difference seems unimaginably great, what sense does it make to adopt a method of utility comparison that treats the two equally? In brief, the objections to the zero-one rule are the same as those which apply to the method of counting perceptual jumps. It certainly provides a way of comparing utils across individuals, but who is to say that the comparisons generated are relevant to anything of interest?

5 Harsanyi and Interpersonal Comparison.

Harsanyi [15] builds on the orthodox Von Neumann and Morgenstern theory of utility. The assumptions of this theory have no bearing on interpersonal comparison, and so he necessarily addds extra assumptions to those of Von Neumann and Morgenstern. It is important to recognize the necessity of making such assumptions, since Von Neumann and Morgenstern's use of "transferable utility" in discussing coalition formation in the second half of their book has left many commentators with the mistaken belief that pure Von Neumann and Morgenstern utility theory allows different individuals, not only to compare their utils, but to pass them from hand to hand if they feel so inclined. My own view is that transferable utility can only make proper sense in the case when the players are all risk neutral, so that their utils can be identified with dollars, but the immediate point is simply that Von Neumann and Morgenstern did not attempt the impossible task of deducing their ideas on transferable utility from the assumptions they wrote into Von Neumann and Morgenstern utility theory.

Harsanyi's [15] additional assumptions are built on the notion of what I call empathetic preferences. Such preferences were introduced by the philosopher Patrick Suppes [32], and studied by Sen [29] and Arrow [1] under the name of "extended sympathy preferences".

Sympathetic preferences. In surveying the history of utilitarianism, Russell Hardin [13] dismisses Hume's emphasis on the importance of sympathetic identification between human beings as idiosyncratic. Although Adam Smith [31] followed his teacher in making human sympathy a major plank in his *Theory of Moral Sentiments*, Hardin is doubtless broadly right in judging that later moral philosophers appeal to human sympathy only when in need of some auxiliary support for a conclusion to which they were led largely by other considerations. Nor is it hard to see why Hume's ideas on human sympathy should have been eclipsed by more peripheral notions. The reasons are much the same as those that led to the eclipse of his even more significant insight into the importance of conventions in human societies. In brief, until game theory came along, no tools were available to operationalize Hume's ideas.

The credit for seeing the relevance of game theory to Hume's idea of a convention probably belongs largely to Thomas Schelling [28]. In the case of Hume's notion of human sympathy, my guess is that it is Harsanyi [15] who saw the way ahead most clearly. In any case, it is Harsanyi's development of the idea that is followed here.

It is first necessary to recognize that what Hume and Adam Smith called sympathy is nowadays known as *empathy;* psychologists reserve the word *sympathy* for a stronger notion. Adam sympathizes or empathizes with Eve when he imagines himself in her shoes in order to see things from her point of view. When Adam sympathizes with Eve, he identifies with her so strongly that he is unable to separate his interests from hers. For example, before the creation of Eve, Adam perhaps took no interest at all in apples while gathering his daily supply of fig leaves. But, after falling in love with the newly created Eve and observing her fondness for apples, he might then have found himself unable to pass an apple tree without salivating at the thought of how much Eve would enjoy its fruit. In such a case, it would not even be very remarkable if he were to abandon foraging for fig leaves altogether, so as to devote himself entirely to gathering apples for her.

The theory of revealed preference has no difficulty in describing Adam's behavior in such a case of sympathetic identification. If Adam chooses to gather apples rather than fig leaves, he reveals a preference for the consumption of apples by Eve to the consumption of fig leaves by himself. If he is consistent in this behavior, it can be described using a Von Neumann and Morgenstern utility function that depends both on his own consumption and on Eve's consumption. No theoretical difficulty therefore exists in incorporating altruistic (or spiteful) preferences into a player's utility function. If Adam really cares for Eve to the extent that he is willing to sacrifice his own physical well-being for hers, then this will be the right and proper way to model the situation.

It is easy to see why the forces of biological evolution might lead to our behaving as though we were equipped with sympathetic preferences. Mothers commonly care for their children more than they do for themselves—just as predicted by the model that sees us merely as machines that our genes use to reproduce themselves. In such basic matters as these, it seems that we differ little from crocodiles or spiders. However, humans do not sympathize only with their children; it uncontroversial that

they also sympathize to varying degrees, with their husbands and wives, with their extended families, with their friends and neighbors, and with their sect or tribe.

Modern behavioral economists are willing to proceed as though we all sympathize with everybody in much the same way that we sympathize with our near and dear. If this were true, then the Von Neumann and Morgenstern theory would be adequate all by itself to determine a standard of interpersonal comparison, because Adam would only need to consult his own sympathetic utility function to find out how many utils to assign to a change in Eve's situation as compared with some change in his own situation. But Harsanyi's [15] approach is less naive. He argues that, alongside our personal preferences (which may or may not include sympathetic concerns for others), we also have empathetic preferences that reflect our ethical concerns.

Empathetic preferences. When Adam *empathizes* with Eve, he does not identify with her so closely that he ceases to separate his own preferences from hers. We weep, for example, with Romeo when he believes Juliet to be dead. We understand why he takes his own life—but we feel no particular inclination to join him in the act. Similarly, a confidence trickster is unlikely to sympathize with his victims, but he will be very much more effective at extracting money from them if is able to put himself in their shoes with a view to predicting how they will respond to his overtures. I think we unconsciously carry out such feats of empathetic identification on a routine basis when playing the game of life each day with our fellow citizens.

It seems evident to me that empathetic identification is crucial to the survival of human societies. Without it, we would be unable to find our way to equilibria in the games we play except by slow and clumsy trial-and-error methods. However, it is not enough for the viability of a human society that we be able to use empathetic identification to recognize the equilibria of commonly occurring games. The games we play often have large numbers of equilibria. As Hume [18] saw so clearly, society therefore needs commonly understood coordinating conventions that select a particular equilibrium when many are available. Sometimes the conventions that have evolved are essentially arbitrary—as in the case of the side of the road on which we drive. However, in circumstances that are more deeply rooted in our social history, we usually overlook the conventional nature of our equilibrium selection criteria. We internalize the criteria so successfully that we fail to notice that selection criteria are in use at all.

We are particularly prone to such sleepwalking when using those conventional rules that we seek to justify by making airy references to "fairness" when asked to explain our behavior. In saying this, I do not have in mind the rhetorical appeals to fairness that typify wage negotiations or debates over taxation. Nor do I have in mind the abstract notions of justice proposed by philosophers like Rawls. I am thinking rather of the give-and-take of ordinary life. Who should wash the dishes tonight? Who ought to buy the next round of drinks? How long is it reasonable to allow a bore to monopolize the conversation over the dinner table? We are largely unconscious of the fairness criteria we use to resolve such questions, but the degree

of consensus that we achieve in so doing is really quite remarkable. My guess is that the real reason the idea of Rawls' original position appeals so strongly to our intuition is simply that, in working through its implications, we recognize that it epitomizes the basic principle that underlies the fairness criteria that have evolved to adjudicate our day-to-day interactions with our fellows.

In order to use Rawls' device of the original position successfully as an equilibrium selection criterion, we need to be able to empathize with other people. In particular, we need to recognize that different people have different tastes. The device would obviously be worthless if Eve were to imagine how it would feel to be Adam without substituting his personal preferences for hers. But more than this is necessary. In order to make fairness judgments, Eve must be able to say *how much* better or worse she feels when identifying with Adam than when identifying with herself. Empathetic identification by itself is not sufficient for this purpose. An essential prerequisite for the use of the original position is that we be equipped with empathetic *preferences*.

Adam's *empathetic* preferences need to be carefully distinguished from the *personal* preferences built into his personal utility function. For example, I am expressing an empathetic preference when I say that I would rather be Eve eating an apple than Adam wearing a fig leaf. My own personal preferences are irrelevant to such an empathetic preference. Since I am no beach boy, I would personally much prefer a fig leaf to cover my nakedness than an apple to add to my waistline. However, if I know that apples taste very sweet to Eve and that Adam is totally unselfconscious about his body, I would clearly be failing to empathize successfully if I were to allow my own impulses towards modesty to influence my judgment about whether Eve is gaining more satisfaction from her apple than Adam is getting from his fig leaf.

Harsanyi's argument. It seems uncontentious that we actually do have empathetic preferences that we reveal when we make "fairness" judgments. Ordinary folk are doubtless less than consistent in the empathetic preferences they reveal, but Harsanyi idealized the situation by taking *homo economicus* as his model of man. In his model, everybody therefore has consistent empathetic preferences, which Harsanyi takes to mean that the Von Neumann and Morgenstern rationality requirements are satisfied. An empathetic preference can therefore be described using a Von Neumann and Morgenstern utility function.

An orthodox personal utility function of the kind we have considered hitherto simply assigns a utility to each situation that the person in question might encounter. For an empathetic utility function we have to pair up each such situation with the person whom we are considering in that situation. One such pair might consists of Adam wearing a figleaf. Another might be Eve eating an apple. An empathetic utility function assigns a utility to each such pair. It is, of course, precisely such pairs of possibilities that must be evaluated when people imagine themselves in the original position behind a veil of ignorance that hypothetically conceals their identity.

The next step in Harsanyi's argument is another idealization. He assumes that

when someone empathizes with Adam or Eve, he does so entirely successfully. More precisely, if I am totally successful in empathizing with Adam, then the preferences I will express when imagining myself in Adam's position will be identical to Adam's own personal preferences. This is an important point. It escapes Parmenio, for example, in the following exchange quoted from Longinus by Hume [18]:

> "Were I Alexander," said Parmenio, "I would accept of these offers made by Darius."
>
> "So would I too," replied Alexander, "were I Parmenio."

Parmenio makes the mistake of putting himself in Alexander's shoes while retaining his own personal preferences. Alexander corrects him by putting himself in Parmenio's shoes with Parmenio's personal preferences.

The rest of Harsanyi's argument is a straightforward application of the properties of Von Neumann and Morgenstern utility functions. The property that matters here is that any two Von Neumann and Morgenstern utility scales that represent exactly the same preferences[11] must be related in the same way as two temperature scales. That is to say, the two scales can differ only in the placing of their zero and their unit. For example, once one knows the number of degrees that the Centigrade and Fahrenheit scales assign to the freezing and boiling points of water, then one knows how to translate any temperature on one scale into the corresponding temperature on the other scale.

The two utility scales to which this fact is now applied are Adam's personal scale and my empathetic scale for Adam. Since Harsanyi's second assumption implies that both scales represent the same preferences, my empathetic scale for Adam is exactly the same as his personal scale except that the zero and the unit are changed. In particular, a util on my empathetic scale for Adam is obtained by multiplying a util on his personal scale by some constant number a. Similarly, a util on my empathetic scale for Eve is obtained by multiplying a util on her personal scale by some constant number e.

It follows that my empathetic utility function simply expresses the fact that I think that his and her personal utils can be traded off at so that e of Adam's personal utils count the same as a of Eve's personal utils. That is to say, Harsanyi's assumptions imply that holding an empathetic preference is exactly the same thing as subscribing to a standard for making interpersonal comparisons between Adam and Eve.

6 Common Interpersonal Comparisons

The interpersonal comparisons described in the preceding section are $idiosyncratic$ to the individual making them. If further assumptions are not made, there is nothing to prevent different people comparing utils across individuals in different ways.

[11] Including preferences over risky alternatives.

Under what circumstances will these different value judgments be the same for everybody in a society? Only then will we have an uncontroversial standard for making interpersonal comparisons available for use in formulating a social contract. Indeed, in the absence of such a *common* standard, many authors would deny that any real basis for interpersonal comparison of utilities exists at all.

Harsanyi [15, p.60] holds that the interpersonal comparisons of utility that we actually make reveal a high degree of agreement across individuals. Rawls [24] agrees with this assessment.But I am not satisfied simply to note the existence of some measure of consensus in the society in which we live. It seems to me that the standard of interpersonal comparison that a society employs is subject to the same forces of social evolution as its social contract. One cannot therefore glibly take a standard for interpersonal comparison as given when discussing social contract issues. One needs to ask how and why such a standard interacts with whatever the current social contract may be—and how it would adapt in response to proposed reforms of the current social contract.

I think that the empathetic preferences with which we find ourselves holding are a product of *social* evolution. We need such empathetic preferences only because they serve as inputs to the equilibrium selection criteria that lead us to speak of "fairness" when we try to explain to ourselves what we are doing when we use them. However, it is important not to allow oneself to be deceived by this propaganda. Our "fairness" criteria do not necessarily treat all citizens in an even-handed mannner—whatever this might mean. As with all social institutions, the "fairness" criteria we use will tend to result in certain types of behavior becoming perceived as more successful than others. Those whose behavior is perceived to be successful are more likely to serve as the locus for meme replication than those who are perceived as failures. The point here is that social evolution will tend to favor the survival of whatever empathetic preferences promote the social success of those that hold them at the expense of those that do not. In the medium run, an equilibrium in empathetic preferences will be achieved. In my books, I argue that, in such an *empathetic equilibrium*, everybody will have the same empathetic preferences and hence we will all share a common standard for making interpersonal comparisons of utility—as Harsanyi and Rawls suggest is actually true for our society. However, this paper is not the place to review my own evolutionary theory.

7 Conclusion

This paper has reviewed only one part of an enormous subject. Its aim has been to clarify how utility is understood by modern economists and to explain why the widespread claims that such a view of utility is incompatible with making interpersonal comparisons is mistaken. It concludes by outlining the theory of interpersonal comparison of John Harsanyi, which I believe is entirely satisfactory for the purpose of making judgments in Rawls' original position.

References

[1] K. Arrow. Extended sympathy and the problem of social choice. *Philosophia*, 7:233–237, 1978.

[2] J. Bentham. An introduction to the principles of morals and legislation. In *Utilitarianism and Other Essays*. Penguin, Harmondsworth, UK, 1987. (Introduction by A. Ryan. Essay first published 1789).

[3] K. Binmore. *Playing Fair: Game Theory and the Social Contract I*. MIT Press, Cambridge, MA, 1994.

[4] K. Binmore. *Just Playing: Game Theory and the Social Contract II*. MIT Press, Cambridge, MA, 1998.

[5] K. Binmore. *Natural Justice*. Oxford University Press, New York, 2005.

[6] G. Cohen. Equality of what? On welfare, goods and capabilities. In M. Nussbaum and A. Sen, editors, *The Quality of Life*. Clarendon Press, Oxford, 1989.

[7] F. Edgeworth. *Mathematical Psychics*. Kegan Paul, London, 1881.

[8] J Elster and J.Roemer. *Interpersonal Comparisons of Well-Being*. Cambridge University Press, Cambridge, 1991.

[9] A. Gibbard. *Wise Choices and Apt Feelings: A Theory of Normative Judgment*. Clarendon Press, Oxford, 1990.

[10] J. Griffin. *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press, Oxford, 1986.

[11] P. Hammond. Why ethical measures of inequality need interpersonal comparisons. *Theory and Decision*, 7:263–274, 1976.

[12] P. Hammond. Interpersonal comparisons of utility: Why and how they are and should be made. In J. Elster and J. Roemer, editors, *Interpersonal Comparisons of Well-Being*. Cambridge University Press, London, 1991.

[13] R. Hardin. *Morality within the Limits of Reason*. University of Chicago Press, Chicago, 1988.

[14] J. Harsanyi. Cardinal welfare, individualistic ethics, and the interpersonal comparison of utility. *Journal of Political Economy*, 63:309–321, 1955.

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

[16] J. Harsanyi. Normative validity and meaning of Von Neumann and Morgenstern utilities. In B. Skyrms, editor, *Studies in Logic and the Foundations of Game Theory: Proceedings of the Ninth International Congress of Logic, Methodology and the Philosophy of Science*. Kluwer, Dordrecht, 1992.

[17] D. Hausman. The impossibility of interpersonal utility comparisons. Technical Report Working Paper DP1/94, LSE Centre for Philosophy of Natural and Social Sciences, 1994.

[18] D. Hume. *A Treatise of Human Nature (Second Edition)*. Clarendon Press, Oxford, 1978. (Edited by L. A. Selby-Bigge. Revised by P. Nidditch. First published 1739).

[19] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

[20] C. Kukathas and P. Pettit. *Rawls: A Theory of Justice and its Critics*. Polity Press with Blackwell, Oxford, 1990.

[21] R. Layard. *Happiness: Lessons from a New Science*. Allen Lane, London, 2005.

[22] R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.

[23] R. Luce and H. Raiffa. *Games and Decisions*. Wiley, New York, 1957.

[24] J. Rawls. *A Theory of Justice*. Oxford University Press, Oxford, 1972.

[25] J. Rawls. Social unity and primary goods. In A. Sen and B. Williams, editors, *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, 1982.

[26] L. Robbins. Inter-personal comparisons of utility. *Economic Journal*, 48:635–641, 1938.

[27] A. Rubinstein. Similarity and decision-making under risk. *Journal of Economic Theory*, 46:145–153, 1988.

[28] T. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

[29] A. Sen. *Collective Choice and Social Welfare*. Holden Day, San Francisco, 1970.

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

[31] A. Smith. *The Theory of Moral Sentiments*. Clarendon Press, Oxford, 1975. (Edited by D. Raphael and A. Macfie. First published 1759).

[32] P. Suppes. Some formal models of grading principles. *Synthèse*, 6:284–306, 1966.

[33] H. Varian. *Microeconomic Analysis (Third Edition)*. Norton, New York, 1992.

[34] J. Von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.