Utility theory from Jeremy Bentham to Daniel Kahneman

Daniel Read

[1] *Department of Operational Research, London School of Economics*

Working Paper No: LSEOR 04-64 ISBN No: 07530 1689 9

First published in Great Britain in 2004
by the Department of Operational Research
London School of Economics and Political Science

Typeset, printed and bound by:
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

Working Paper No: LSEOR 04-64 ISBN No: 07530 1689 9

Utility theory from Jeremy Bentham to Daniel Kahneman

A standard model of motivation is that a person has a desire Y, and if they believe that by doing act X, they can achieve Y, then (assuming there is no barrier to doing X or some stronger desire than Y) they will choose X. The normative problem of rationality concerns what choices and desires people should have. The most well-established approach to this problem is *rational choice theory,* which prescribes the most effective ways to achieve *given* desires (Sugden, 1991). The only constraints rational choice theory puts on desires is that they be consistent. Many observers are dissatisfied with such a purely structural definition of rationality, and want rules of rationality to say something substantive about *what* desires are best. Dennett (1981), for example, voiced this concern when he included as one of his principles of rationality, that 'a system's desires are those it ought to have, given its biological needs and the most practical means of satisfying them.' Less formally, we would like our account of rationality to answer such questions as: is it rational for a smoker to smoke, for an obese person to overeat, or for an employee to undersave? While we might think not, rational choice theory does not put constraints on what these people *ought* to want.

The utilitarian philosophers, including Bentham, Mill and Sidgwick, did put forward substantive rules that can be treated as rules of rationality. Roughly, they proposed that people ought to desire those things that will maximise their *utility,* where positive utility is defined as the tendency to bring pleasure, and negative utility is defined as the tendency to bring pain. This utilitarian viewpoint, especially as advocated by Jeremy Bentham, had a significant impact on 19th century economics. But Bentham's utilitarian project was eventually abandoned in favour of structural accounts of rationality and formal definitions of utility such as rational choice theory. A major reason for this abandonment was that Benthamite utility was deemed impossible to measure. Since Bentham's time, however, the social sciences have developed greatly, and armed with more sophisticated methods, Daniel Kahneman and his co-workers have proposed that we go 'Back to Bentham' (e.g., Kahneman, Wakker & Sarin, 1997)[i]. The result is an economic psychology based on the measurement of *experienced utility.* Kahneman's ambitious program is in its early stages, but if successful, it promises to alter our understanding of rationality, by allowing us to assert not only that X is the rational way to achieve Y, but also that Y the rational thing to want to achieve. Or, that not only is the smoker annoying, he is also irrational.

The purpose of this paper is to sketch out the history of the idea of utility in its circuitous path from Bentham to Kahneman, and then to consider the problems and challenges that still remain. We begin with the history.

A selective history of utility theory[ii]

Jeremy Bentham's (1748-1832) moral philosophy centred on two assumptions: the goodness or badness of experience is quantifiable, and the quantities so obtained can be added across people. The first assumption is exemplified by the following famous passage:

> To a person considered by himself, the value of a pleasure or pain considered by itself, will be greater or less, according to the four following circumstances: 1. Its *intensity*; 2. Its *duration*; 3. Its certainty or uncertainty; 4. Its propinquity or remoteness. (IV, 4, italics added).

In short, if we multiply, for that person, the intensity of feeling times its duration we obtain its value for that person. The second assumption is that we can add up the individual degrees of value to get a measure of the social good. Bentham continues:

> Take an account the number of persons whose interest appear to be concerned, and repeat the above process with respect to each. *Sum up* the numbers … . Take the *balance*; which, if it be on the side of pleasure, will give the general *good tendency* of the act, with respect to the total number or community of individuals concerned…. (IV, 6, Bentham's italics).

And so on. The goal of an action (or, as we might put it, the good) was to maximise utility, or, as it was put succinctly by Edgeworth (1879): "The greatest possible value of ∫∫∫*dn dt dp* (where *dp* corresponds a just perceivable increment of pleasure, *dt* to an instant of time, *dn* to a sentient individual.)"

As used by Bentham and his followers, *utility* was the tendency of an object or action to increase or decrease overall happiness. Benthamite utility is also logically separate from what choices are actually made: Someone, such as a smoker, might choose something that has lower utility than its alternative. Therefore, simply knowing what people want will not tell us what they should have.

Bentham's ideas had a profound influence on the economists of his and subsequent generations, most notably such classical economists such as Gossen (1810-1858), Jevons (1835-1882), Marshall (1842-1924) and Edgeworth (1845-1926). Jevons (1888), in the preface to *The Theory of Political Economy*, stated that "In this work I have attempted to treat Economy as a Calculus of Pleasure and Pain. (PF. 3)." These economists looked forward to a time when utility could be measured directly (Edgeworth proposed a 'hedonimeter') although, they also believed that the best approximation they had was behaviour in the marketplace:

> A unit of pleasure or pain is difficult even to conceive; but it is the amount of these feelings which is continually prompting us to buying and selling, borrowing and lending, labouring and resting, producing and consuming; and it is from the quantitative effects of the feelings that we must estimate their comparative amounts. (I. 17).

As hinted by this passage, while pleasures and pains constituted the metaphysical foundation of utilitarian economics, neither their measurement nor even their existence was central to their *methods*. Rather, choice behavior was assumed to reflect, however roughly, the quantity of utility derived from a choice. Marshall (1920), indeed, was explicit about what assumptions were being made, and their potential shortcomings (even as he distinguished between what we will later be calling decision utility and experienced utility):

> It cannot be too much insisted that to measure directly, or *per se,* either desires or the satisfaction which results from their fulfilment is impossible, if not inconceivable. If we could, we should have two accounts to make up, one of desires, and the other of realized satisfactions. And the two might differ considerably. … But as neither of them is possible, we fall back on the measurement which economics supplies, of the motive or moving force to action: and we make it serve, with all its faults, *both* for the desires which prompt activities and for the satisfactions that result from them. (Book III, Chapter III, Footnote 57).

Likewise, while the concept of *total utility*, meaning the total pleasure or pain that choices brought, was central to normative economic thinking, only *marginal* *utility*, meaning the pleasure or pain from an additional unit or 'dose' of a good was needed in their economic analysis. The marginalist revolution – due in large part to the economists cited above – based its account of individual decision making on what Stigler (1950a) calls the 'fundamental principle of marginal utility theory:' In Gossen's words, 'Man maximizes his total life pleasure if he distributes his entire money income … among the various enjoyments … so that the last atom of money spent on each single pleasure yields the same amount of pleasure (cited in Georgescu-Roegen, 1968, p. 244).' In symbols, this condition is:

$$\frac{MU_1}{p_1} = \frac{MU_2}{p_2} = ... = \frac{MU_i}{p_i}, \text{ for all } i.$$

Where MU_i is the marginal utility of good i, and p_i is its price. The Benthamite project, therefore, was further reduced because there was no longer any need to measure, or even to theorize about, total pleasure or pain. Even if marginal utilities were measured on a cardinal scale, they would tell us nothing about how much total utility there was (even if it was maximized) because they are still only measured up to an additive constant. As Georgescu-Roegen puts it:

> The level of utility can be visualized as a bottomless ocean; the wave on top can, nevertheless, be seen by a navigator and described by the curve of marginal utility. (p. 241).

But if we want to know the *volume* of the ocean, knowing about the waves is of very little help.

Marginal-utility nonetheless retained a degree of cardinal measurability in that differences between increments could be interpreted on a ratio scale: it was still possible to state that the difference between a cup of coffee and a cup of tea was greater than that between a cup of tea and a glass of lemonade. A new analytical problem arose, however, that eventually undermined even marginal utility. This was the fact that utilities are not independent – the marginal utility from a kilogram of coffee, for instance, depends on whether one owns an espresso machine or only a saucepan. Edgeworth dealt with this problem by proposing that total utility is a function of the entire basket of goods or opportunities faced by the consumer, or $U=U(x_1,x_2,\ldots,x_n)$. The utility of a bundle of goods could be represented in a multi-dimensional space (one dimension per good) with more preferred bundles located above less-preferred ones. Bundles having the same value could be joined together to form an indifference curve. This analysis soon led to the ordinal revolution in utility theory, which eliminated all reference to total utility. Thus, although Edgeworth was a devoted Benthamite, 'by an irony of history, the ardent utilitarian thus became the pathfinder of ordinalism. (Alchian, 1990, p.282).'

Pareto (1848-1923) rejected altogether the idea that quantities of utility mattered. He observed that if we map preferences onto Edgeworth's indifference curves, we know everything necessary for economic analysis. To map these preferences, we obtain pairwise comparisons between possible consumption bundles. The agent will either be indifferent between each bundle, or else will prefer one to the other. By obtaining comparisons between all bundles, we can draw a complete map of an individual's utility. To predict his or her choices under a given budget constraint, we then need only to determine which bundle(s) are on the highest achievable indifference curve. Conversely, we can also assume, from the individual's choice, that because a bundle is chosen, it must be on that highest indifference curve (the principle of *revealed preference*). Observe that Pareto's procedure makes no reference to *any* cardinal utility measure. If you assign a number to each indifference curve, the set of numbers you assign will not matter as long as higher indifference curves get higher numbers. Any monotonic increasing transformation (such as multiplication by a positive constant) can be applied to those numbers and all the necessary information will be retained. This *ordinal* utility function, therefore, represents a complete abandonment of Benthamite utilitarianism. First, utility no longer has any relationship with degrees-of-happiness, and even the promissory note of Jevons and Marshall has been abandoned. Second, the numbers assigned to bundles cannot be combined across people. Third, even the differences between these numbers are incomparable. Moreover, there is no suggestion (as was found in the passages from Jevons and Marshall) that ordinal utility is merely a half-way house to eventual cardinality once the science of measurement has developed sufficiently. Hicks (1939), the arch-ordinalist, underlined this when he observed that:

> … this does not mean that if any one has any other ground for supposing that there exists some suitable quantitative measure of utility, or satisfaction, or desiredness, there is anything in the [ordinal-utility] argument to set against it. If one is a utilitarian in philosophy, one has a perfect right to be a utilitarian in one's economics. But if one is not (and few people are utilitarians nowadays), one also has the right to an economics free of utilitarian assumptions. (p. 18).

An apparent partial return to cardinal utility was brought in by a new approach to utility measurement, first proposed by Von Neumann and Morgenstern. They showed that just as an indifference map can be drawn from consistent choices between outcomes, so a cardinal utility function can be drawn from consistent choices between gambles. Consistency, in this case, means that choices conform to a set of axioms that are logically equivalent to a utility function. The axioms are simple principles which all reasonable people will agree should be conformed to by a rational decision maker. Because of the logical equivalence, if the agent's preferences conform to these axioms, then they can be described with a utility function that is unique up to a linear transformation. There is no necessary relationship between utility, as derived from consistency with the axioms, and the experience of the decision maker, or his satisfaction (see, Ellsberg, 1954). The utility function summarizes the person's preferences and nothing more. So the return of measurable utility offered no comfort to the Benthamite.

Back to Bentham: Experienced utility

The preceding section showed how economics became increasingly separated from the Benthamite goal of maximising the social good. This goal, however, has been revived by a new empirical approach to utilitarianism, based in large part on Daniel Kahneman's theory of experienced utility[iii]. This theory is based on the belief that that there is a 'measurable' good that is separable from the choices people make. In this section, I begin with a brief introduction to experienced utility, which then leads to a discussion of what needs to be done for this theory to achieve its Benthamite ambitions.

Before a cautionary remark is in order. Kahneman is very aware of many problems with the application of the theory of experienced utility, and thus reluctant to make broad generalisations. We can distinguish, therefore, between a strong and a weak version of the theory. The strong version, the one closest to Bentham, is in my view the one with the greatest potential, but one which Kahneman should not be required to justify. Kahneman, Wakker and Sarin (1997), for example, explicitly distance themselves from the strong view by sharply limiting its application:

> Our normative treatment of the utility of temporally extended outcomes adopts a hedonic interpretation of utility, but no endorsement of Bentham's view of

> pleasure and pain as sovereign masters of human action is intended. Our analysis applies to situations in which a separate value judgment designates experienced utility a criterion for evaluating outcomes.' (p. 377).

Although most of the comments that follow apply to any version of the theory of experienced utility, I direct my attention to the strong theory.

The premises of the (strong) theory of experienced utility can be stated in a few propositions. Firstly, at every moment we are experiencing utility, meaning pleasure and/or pain (this is termed *instant utility*). Second, this utility has *quantity* and *valence*, with a neutral point on the boundary between desirable and undesirable, pleasure and pain. Third, keeping in mind Kahneman's own reservations, this utility is all that makes an experience good or bad. Fourth, by integrating instant utility over a period we obtain the *total utility* for that period. Fifth, an optimal decision is one that maximises total utility (or expected total utility). Finally, to make this a workable theory, instant utility must be measurable, up to at least an ordinal and ultimately a ratio scale[iv].

Kahneman and co-authors distinguish between experienced utility and *decision utility*, which is the utility reflected in choices, or revealed preferences. Kahneman emphasizes there is no logical requirement for experienced utility and decision utility to coincide, and that if the two utilities differ in their prescription, we may want to favour experienced utility. This fits our intuitions about smokers: they may decide to have a cigarette (decision utility), yet be better off if they don't (experienced utility).

Kahneman and his co-workers have conducted many studies demonstrating how experienced and decision utility can differ. These studies draw on a normative principle of rationality, *temporal monotonicity*, analogous to the principle of stochastic dominance in risk. Temporal monotonicity means that if you have two identical experience sequences, and improve part of one sequence, then the changed sequence should be preferred to the unchanged one, while if you worsen part of it, the unchanged sequence should be preferred:

> Consider two episodes that are preceded and followed by a steady state of hedonic neutrality. Assume that the second episode is obtained by adding an unanticipated period of pain (or pleasure) to the first, prior to the return to the neutral state. The monotonicity principle asserts that the hedonic quality of the added period determines whether the longer episode has higher or lower global utility than the shorter. In other words, adding pain at the end of an episode must make it worse. Adding pleasure must make it better.

Kahneman et al. (1993) have conducted several tests to determine if decision utility conforms to this principle. In one, participants first experienced two painful immersions of a hand in cold water. Both hands were immersed (at different times) in ice-cold (14.1°C) water for 60 seconds. One hand was then removed immediately, while the other was left in the water, which warmed up to a still uncomfortable 15.2°C over an additional 30 seconds. The *total* pain, meaning the sum of instantaneous pain multiplied by duration, was certainly greater for the second hand, yet when asked which of the two experiences they would prefer to repeat almost 70% chose the longer one. This shows that decision utility differs from experienced utility. Moreover, decision utility appears to be mistaken because it directs the decision maker to suffer unnecessary pain.

Kahneman also considers why decision utility systematically differs from experienced utility. One reason is that *remembered* utility, presumably a major determinant of decision utility, is a gestalt impression constructed from only a small and often biased sample of experience. He proposes the peak-end rule, which is that remembered utility is the average of the instant utility measured at the peak and the end of an experience. In the cold water study, for instance, the peak pain in the two experiences was identical, while the pain at the end (14.1°C versus 15.2°C) was not. When the two quantities are averaged in memory, the longer experience is preferred, violating temporal monotonicity.

Such studies (and there are many) make it clear that what people will choose often differs from what gives them the most pleasant (or least unpleasant) experiences. Moreover, they suggest it is possible to objectively measure the pleasantness and unpleasantness. Admittedly, up to now the research has focused on simple events involving simple sources of pleasure and (mostly) pain, but it offers hope that Bentham's utilitarian project can be successfully revived.

In the next three sections I consider how the theory of experienced utility has dealt with three challenges that have dogged utilitarian theory. The first concerns dependent utilities: the fact that the utility of an act depends not only on the person's tastes, but also on everything else that person has or does. The second concerns the problem of the indeterminate origin of the utility function (the ocean problem). The third is the ultimate test for any utility theory that has welfare implications –whether one person's utility can be compared with another's. As will become clear, I believe only some of these challenges have been met successfully.

Dependent utilities

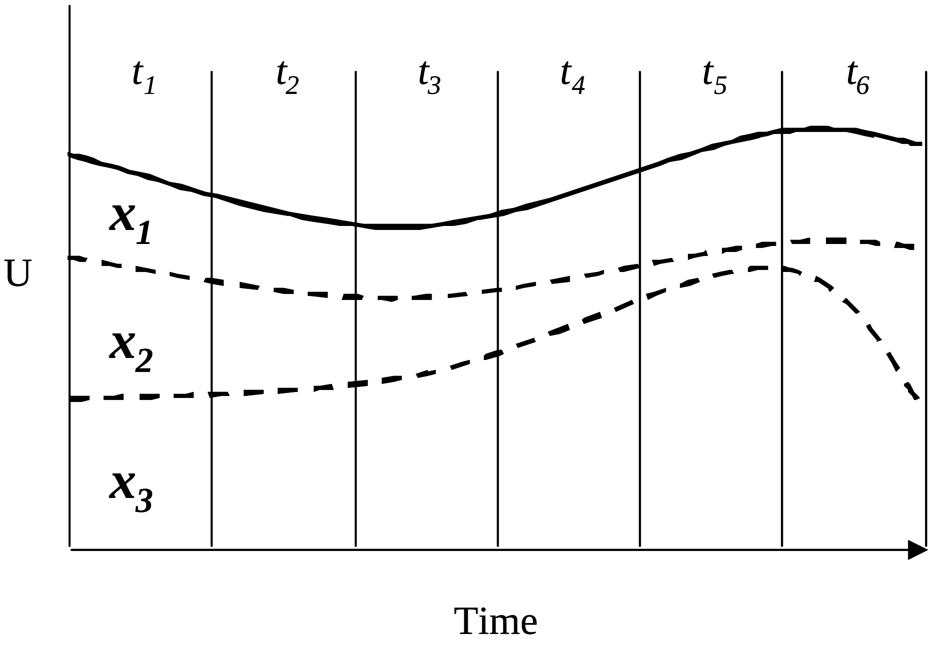
As already discussed, early economists (i.e., pre-Edgeworth) assumed, at least for analytic purposes, that the utility from a given quantity of goods was independent of what other goods were possessed. That is, the total utility of the bundle $(x_1,x_2,\ldots, x_i)$ is given by $u(x_1)+u(x_2)+\ldots+ u(x_2)$. This is an unrealistic assumption, as can be readily seen. For instance, the utility contribution of each member of a complementary pair, like a loaf of bread and a toaster, is the sum of their utilities conditionalised on the other item in the pair: u(toaster, bread) = u(toaster|bread) + u(bread|toaster). In reality, there are no goods that are completely independent, so the total utility has to be calculated as:

$$\sum_i u(x_i \mid \{x_j, j \neq i\}).$$

From the measurement perspective, calculating total utility in this way is practically impossible. Moreover, the consumption bundle is itself an idealization. No one has ever, nor ever will have, constructed an account of a single consumer's consumption bundle. To attempt to do so would be the economic analogue of the genome project, except that while we can unambiguously identify a gene, we cannot unambiguously identify a good.

Both these problems are overcome by the measurement of experienced utility. Instant utility is the aggregate effect of the entire consumption bundle as it is experienced at every moment. This is illustrated in Figure 1. Imagine that the individual has a consumption bundle of three goods, and that the stacked line chart gives the contribution of each good to current happiness (the utility contribution is the difference between one line and the next). To know instant utility, all we need to know is how the person feels at a given time. We do not need to know what contributes to that feeling, nor by how much. Moreover, if we want to know how much a specific good will contribute when added to a person's consumption bundle, we only need to measure instant utility up to the point when the good arrives, and then after it arrives.

The theory of experienced utility also deals with the closely related issue of *temporal separability.* It is well known that our experience at one time can influence our experience at another time, with great experiences making good ones look bad, and bad experiences making good ones look great (Frederick & Loewenstein, 1999). The value of a sequence, therefore, cannot be obtained by summing the separate values of each outcome in a sequence, without knowing their value in their temporal context. Measurements of instant utility, however, *can* be summed (or *concatenated*) without knowing their temporal context, because these measurements are made in their temporal context, and thus automatically take it into account. This means that while one who receives £30 today and £60 tomorrow might be happier than one who receives £60 today and £30 tomorrow, yet one who experiences 30 units of utility today and 60 tomorrow is no happier than one who experiences 60 units today and 30 tomorrow.



In this way, the theory of experienced utility is able to overcome a major barrier to the Benthamite project by measuring total utility despite the impossibility of determining the contribution of each action to their utility. As we discuss next, however, two further barriers to that project are not yet fully resolved.

The origin of the utility function

To measure total utility it is essential to identify a true zero-point relative to which utility measurements can be made. Georgescu-Roegen's (1968) 'ocean problem' arose because the zero-point was the bottom of the ocean. This implies that the point of 'zero utility' is the worst possible state. Kahneman, however, like Bentham before him, locates the zero point as a neutral state between bad and good. The problem is then changed to the much simpler one of determining the average height of a mountain range, where sea-level is zero.

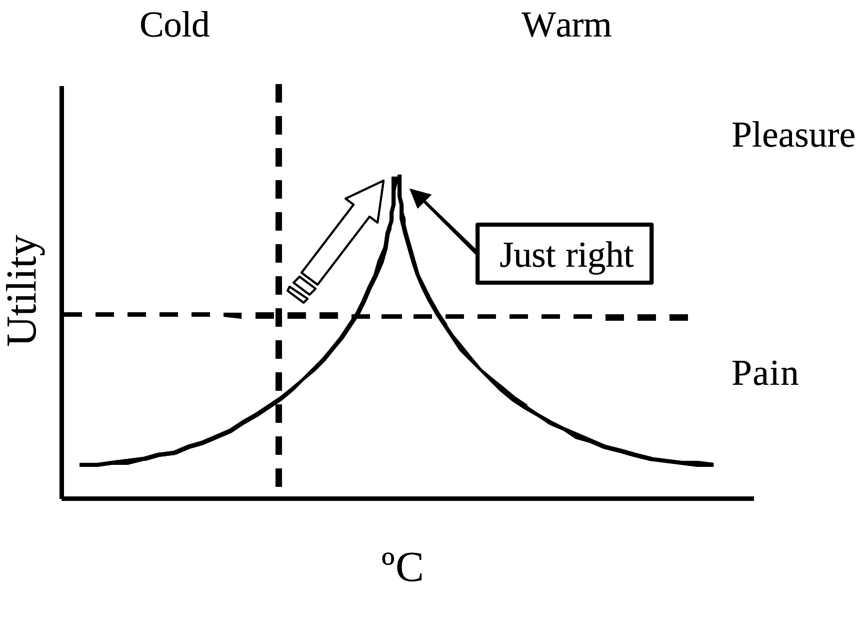
The zero point is defined in two ways: First, as a state where experiences are neither pleasant nor unpleasant; second as the dividing line between stop/go signals for the decision maker:

> The natural zero of the scale of moment utility should be 'neither pleasant nor unpleasant – neither approach nor avoid.' (Kahneman, 2000a, p. 684)

> Pleasure is evidently a "go" signal, which guides the organism to continue important activities such as foreplay or consuming sweet, energy-rich food. Pain is a "stop" signal, which interrupts activities that are causing harm, such as placing weight on a wounded foot. (Kahneman, Wakker & Sarin, 1997, p. 379).
> Like hot and cold, the experiences of pleasure and distress differ in quality. A scale that ranges from extreme pleasure to extreme distress (or from very hot to very cold) effectively comprises separate scales for two distinct attributes. The two scales are joined by a distinctive neutral point. … *The stimulus that gives rise to a neutral experience may be different in different contexts, but the neutral experience itself is constant*. (italics added, Kahneman, Wakker & Sarin, 1997, p. 380).

A key feature of the neutral state is that it is a point where *experience quality* is invariant. The italicized sentence concluding the above passages emphasizes that sometimes it requires a higher or lower temperature to induce a neutral feeling between 'hot' and 'cold,' but that the feeling itself doesn't change. Kahneman's view is that there is a similar neutral state between pleasure and pain.

I admit to doubts about whether such a neutral hedonic state actually exists. To illustrate these, I will consider the pleasure and pain, and not the sensation of hot and cold, caused by heat. This is illustrated in Figure 2. The temperature is denoted by the x-axis, and the feeling associated with that temperature is denoted by the y-axis. The vertical dotted line depicts the division between hot-and-cold (the neutral state), and the horizontal line the division between pleasure and pain. There is also an optimal level of temperature, denoted 'just right' in the figure. Can we specify the two neutral points, the one where cold becomes warm, the other where unpleasant becomes pleasant? We have all done the experiment many times. Imagine you have filled the bathtub with hot water. You put a toe in, but find it is too hot. Nonetheless, you force yourself to get in. The bath cools, and shortly you find it is too cold, and you either have to add hot water or get out. At some point you crossed the neutral point between hot and cold. But you didn't cross a neutral point between pleasant and unpleasant. Rather, the bath was mildly unpleasant when you got in, and mildly unpleasant when you got out. Moreover, the point at which you crossed from warm to cold, and the point when the bath moved from being unpleasantly cold to unpleasantly warm was the same point – it was, in fact, the point labeled 'just right' in the figure.



There was no identifiable neutral point in the bath, therefore, but an *optimal* one. As far as temperature is concerned it could not get any better. We can make a similar observation for hunger. When you are hungry you eat; you reach a state of satisfaction and do not want any more; this is not a neutral state, but the best possible one. If you eat too much or too little you are dissatisfied. Moreover, there is no state corresponding to 'neither hungry nor not-hungry.'

To return to the bathtub again. Imagine you get into a too-hot bath with a good book. Your desire to read the book is strong, and the bath is cooling. Because you are gaining some pleasure from reading, however, you stay in the bathtub until the coldness of the water outweighs the reading pleasure. The better the book, the longer you will remain. The decision to get out of the bath is a manifestation of the 'stop' signal in action, and although we could treat the sensation at that point as corresponding to the zero point, it will not meet the requirement of being context invariant. The decision depends not only on what we are doing at the time, but also on what the options are. If we plan to go straight to bed and continue reading after getting out of the bath, then the stop signal will come much earlier,and when we are experiencing a higher level of pleasure, than if we have to get out and complete our tax return. The stop and go signal always depends on *what* the options are. That is, we choose A over B because A is better than B, not because A is pleasurable and B is painful. I do not think there is any way of measuring, or even conceptualising, the zero-point in the absence of specific choice options. This means that the stop-go point itself has no meaning outside of the choice context, and does not depend on the presence of a zero-point[v].

Even if we cannot define a point of neutral sensation, however, this does not mean we cannot isolate times when we are experiencing something that can be called 'pleasure' and something that cannot. For instance, when discussing what it is to be happy he proposes that:

> As a first approximation, it makes sense to call Helen 'objectively happy' if she spent most of her time in March engaged in activities that she would rather have continued than stopped, little time in situations she wished to escape, and … not too much time in a neutral state in which she would not care either way. (Kahneman, 1999, p. 7).

This first-approximation is reminiscent of Gilbert Ryle's (1949) view of pleasure:

> '… to enjoy doing something, to want to do it and not to want anything else are different ways of phrasing the same thing. … To say that a person has been enjoying digging [in the garden] is not to say that he has been both digging and doing or experiencing something else as a concomitant or effect of the digging… his digging was the pleasure, and not a vehicle of his pleasure.' (p. 108).

Putting aside the issue of a neutral state, both Ryle and Kahneman point out there are times when we wish we were doing something else, and times when we do not. But we don't wish we were doing something else merely because we are in a positively unpleasant state, but because we have something better to do. We can be enjoying ourselves (i.e., be above the neutral point) yet still dissatisfied.

Interpersonal comparability

One of the major shortcomings of most approaches to measuring utility is that the measurements cannot be used as Bentham wished – to measure the total welfare implications of an act. The central problem has always been that of interpersonal comparability. We have no warrant for taking measurements of utility from different individuals and combining them into meaningful aggregates. Kahneman (2000a) suggests that experienced utility solves, or at least has the potential to solve, this problem. The argument is based on three empirical observations. Firstly, there is often strong interpersonal agreement about the effects of experiences, especially of pain. Kahneman cites Algom and Lubel's (1994) finding that 'the relation between a measure of the physical strength of labor contractions during childbirth and self-reports of pain was generally similar for different women' (Kahneman, 2000a, p. 684). Second, there is strong agreement in experience ratings between actors and observers: By observing someone's winces and groans, it is possible to gain a pretty good idea of the strength of their feeling. Finally, there is a strong correlation between self-reported pleasure and pain and physiological measures (e.g., Davidson, 2000).

I suggest these observations do not, however, really address the *core* problem of interpersonal comparability. There are two varieties of comparability. The first is whether we can say, at least to a first approximation, that certain experiences are universally worse or better than others; the second is whether the magnitude of the experience for one person is comparable to that of another. Although the second problem is the one that must be solved before we can assess total social utility, only the first is addressed by the evidence provided. We should not be surprised that many stimuli (temperature, pressure, the death of a loved one) yield feelings that can be predictably ranked. We also should not be surprised that there are certain hedonic universals, such as satiation and habituation. But this does not really solve the problem of interpersonal comparability, which is whether the units on the measurement scale correspond to the same level of sensation for everyone, and not whether experience A is always worse than B. To illustrate this, suppose several people rate the pain from 10 shocks of different intensity. Virtually everyone's ratings will follow the objective intensity of the shock, and it is even likely that a similar Fechnerian formula will fit shock intensity to pain intensity for each person. Yet this does not mean the shocks *felt* the same to each person.

Although the kind of interpersonal comparability actually achieved does allow for some judgments about which society is better, these judgments may be no different than those that can be made in the absence of this comparability. They are, in fact, judgments of Pareto optimality. Imagine that we could measure each person's total utility under two different social conditions. Because we assume that for each person, a higher total utility means a better society, we do know that if in Society A everyone has at least as much total utility as in Society B, but at least one person has more utility in A then in B, then we should choose A. But if there are two distributions in conflict, then even knowing everyone's total utility we cannot adjudicate between outcomes. If I am better off in A than in B, and even if everyone else is better off in B than in A, it is not certain that the optimal *total* distribution is B. Rather, we have to fall back on the same intuitions or rules-of-thumb that economists and philosophers have always used to overcome the problem of interpersonal comparability. For instance, we might assume that everyone has the same utility function, that the better off benefit less from an additional unit of good than the less well off, or that there are certain needs more fundamental than others.

I have suggested that the theory of experienced utility does not yet resolve all problems posed by the challenges that have been directed at Benthamite notions of utility. But the empirical work that has been guided by the theory, as well as that which supports the theory, nonetheless does go a considerable way to resolving the problems. Moreover, the fact that the questions are now explicitly part of empirical science and not, as they were to Bentham and his followers, virtual pipe-dreams, means that the problems are very likely to be solved. In the next section I turn to a different kind of question. Not whether we can reliably measure experienced utility at a level that permits interpersonal comparison, but if, when we can so measure it, we should judge experienced utility or decision utility to be the better measure of what is good.

Experienced and decision utility and the good life

As has been emphasized throughout this paper, the goal of utilitarianism, as well as the 'strong' theory of experienced utility, is to find an objective index of 'the good,' meaning that which makes actions objectively better or worse. This involves at least two criteria: (a) the candidate index must be measured or assessed in some way that can guide decision making, and (b) the index so measured must be the good. Both decision utility and experienced utility meet the first criterion. Decision utility is revealed through choice, and experienced utility is measured through psychophysical methods. Any controversy, therefore, turns on the second criterion. We need to determine on what basis a utility measure should be judged to measure the good.

The Benthamite view is that the correct basis is whether it maximizes pleasure or minimizes pain. There is clear evidence that decision utility does not meet this criterion, and therefore fails this construal of *personal* utility maximisation[vi]. If, therefore, Benthamite utility can be identified with the good, then maximizing experienced utility is a better decision rule. But the question remains whether Benthamite utility can be identified with the good. Or, more precisely (since we are choosing between principles) does experienced utility give a *better* approximation to the good than does decision utility? The answer to this question cannot be automatically answered in the affirmative. Kahneman, Wakker & Sarin (1997) themselves, as we saw in a passage cited above, agreed that 'a separate value judgment' is necessary before Benthamite utility can be identified with the good.

But if we loosen the connection between measured experienced utility and the good, even the bedrock principle of temporal monotonicity can be called into question. To see why, let us consider the apparently straightforward case of whether it is always better to have your hand in cold water for a shorter period. If we start by taking the Benthamite view that pain is a bad in itself, and since longer exposures mean more pain, the short exposure is always better.

But we might also think that it is not pain that is bad, but what pain tells us. Pain, in this view, is merely a signal for us to stop doing what causes pain, and so it is not the pain we should stop, but something else that happens to be associated with pain. If we put our hands in the fire, for example, our subsequent chances of survival are enhanced by the act of removing our hand from the fire, which just happens to be the same as the act of avoiding pain. But if a painful experience has been demonstrated to have no consequences, the signalling value of pain, *in that context*, has lost some of its informativeness, and our evaluation of the signal should be correspondingly updated. If the value of pain is that it is a signal, then the temporal monotonicity principle may not apply in contexts in which pain has no useful signalling power. In short, if pleasure and pain are not the good but the *signal* of the good, then this opens the door to the possibility that experienced utility can be just as 'mistaken' as decision utility.

So far, however, while this shows that experienced utility is not *necessarily* the best index of the good, it does not allow us to adjudicate between experience and decision utility. Assume that the process of evolution has built us to make choices that maximise future reproductive success, and that pleasure and pain signal actions that, from an evolutionary standpoint, have been successful or unsuccessful. Indeed, this assumption underlies Kahneman, Wakker and Sarin's (1997) observation that '…the effect of natural selection is to increase overall fitness, not necessarily to produce organisms that maximise pleasure and minimise pain over time.'

One way overall fitness can be increased is by taking future feelings into account. If we choose an action likely to maximise future pleasure, we are also choosing the action likely to maximise future fitness. Yet the apparent failures of decision utility show that nature has given us an imperfect mechanism for maximising future pleasure. We do a bad job of predicting future feelings, and we also systematically misremember past feelings. We do not remember the total utility from past events, but rather remember summary statistics and selected moments. We remember the peak and the end, but underestimate or even ignore duration. One interpretation of this is that when we misremember utility, and then make choices based on this erroneous memory, nature has made a mistake.

One possible mistake is to systematically misestimate the fitness consequences of our actions. That is, fitness might be better predicted by experienced utility than by decision utility. I suggest this is unlikely, and that if we follow this line of argument, decision utility has the greater a-priori claim to being the 'correct' valuation of outcomes. Decision utility is the output of a flexible decision making system that we do not fully understand, but we know that it has been honed, through evolution, to make optimal decisions. Experienced utility is but one measure of one input to the decision utility system, and it appears that nature has learned to put only partial reliance on that input. To identify experienced utility with the good, therefore, is to second guess nearly five million years of evolution.

This is not the end of the argument, however, as we are no longer living in the African savannah, and rules that were the best ones during the Paleolithic era may have little or no significance for today's humans (e.g., Kanazawa, 2004). It is possible, for instance, that fitness in today's environment can be better approximated by total experienced utility than by decision utility. Or, perhaps more importantly, perhaps we shouldn't care about evolutionary fitness at all. According to this view, we should set as our goal the maximization of experienced utility regardless of whether it serves any additional function. Experienced utility is an *end in itself*, and we should not be slaves to our natural utility function. We can reject nature's strictures, and judge pleasure and pain (or, more generally, happiness) to be the ultimate good after all. This truly does bring us 'Back to Bentham.'

References

Bentham, J. (1824/1987). An introduction to the principles of morals and legislation. In J. S. Mill and J. Bentham, *Utilitarianism and Other Essays*, Harmandsworth: Penguin.

Davidson, R.J. (2000) 'Affective style, psychopathology and resilience: Brain mechanisms and plasticity', *American Psychologist*, 55, 1196-1214.

Davidson, R.J., Jackson, D.C. and Kalin, N.H. (2000) 'Emotion, plasticity, context and regulation: Perspectives from affective neuroscience', *Psychological Bulletin*, 126, 890-906.

Edgeworth, F. Y. (1879) The hedonical calculus. *Mind*, 4, 394-408.

Frederick, S. & Loewenstein, G. (1999). Hedonic adaptation. In Kahneman, D., Diener, E., & Schwarz, N. (Eds.) *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*, 45-55.

Georgescu-Roegen, N. (1973). *Utility and value in economic thought*. In P. Weiner (Ed.). Dictionary of the History of Ideas: Studies of Selected Pivotal Ideas. New York: Scribners.

Georgescu–Roegen, N. (1968) Utility, *International Encyclopedia of the Social Science*s, Vol. 16 . D. L. Sills (Ed). New York: Macmillan and The Free Press.

Jevons, William Stanley. (1888) *The Theory of Political Economy*. London: Macmillan & Co. (Online, Library of Economics and Liberty. http://www.econlib.org/library/YPDBooks/Jevons/jvnPE1.html)

Hicks, J. R. (1939). *Value and Capital: An inquiry into some fundamental principles of economic theory (2d edition)*. Oxford: Clarendon.

Kahneman, D. (1999). Objective happiness. In Kahneman, D., Diener, E., & Schwarz, N. (Eds.) *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.

Kahneman, D. (2000a). Experienced utility and objective happiness: A moment-based approach. In Kahneman, D. & Tversky, A. (Eds.) *Choices, values and frames*. Cambridge: Cambridge University Press.

Kahneman, D. (2000b). Evaluation by moments: Past and future. In Kahneman, D. & Tversky, A. (Eds.) *Choices, values and frames*. Cambridge: Cambridge University Press.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A. & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*, 401-405

Kahneman, D. & Snell, J. Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making*, 1992, *5*, 187-200.

Kahneman, D. and Snell, J. Predicting Utility. In R.M. Hogarth (Ed.) *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. Chicago: University of Chicago Press, 1990, 295-310.

Kahneman, D., Wakker, P. & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics, 112*, 375-406.

Kanazawa, Satoshi. (2004). The Savanna Principle. *Managerial and Decision Economics*. 25

Layard, R. (2003). *Happiness: Has social science a clue?* The Lionel Robbins Lectures 2002/2003.

Marshall, A. (1920). *Principles of Economics*. Macmillan and Co., Ltd. 1920. (Online: Library of Economics and Liberty. www.econlib.org/library/Marshall/marP0.html.)

Mill, J. S. (18/1987). Utilitarianism. In J. S. Mill and J. Bentham, *Utilitarianism and Other Essays*, Harmandsworth: Penguin.

Niehans, J. (1989). *A history of economic theory: Classic contributions*. Johns Hopkins University Press.

Ryle, G. (1949). *The concept of mind.* NY: Barnes and Noble.

Sarin, R., & Wakker, P. (1997). *Benthamite utility for decision making*. Working paper, CenTER, Tilburg, The Netherlands.

Stigler, G. (1950a). The development of utility theory I. *The Journal of Political Economy*, *58*, 307-327.

Stigler, G. (1950b). The development of utility theory II. *The Journal of Political Economy*, *58*, 373-396.

Endnotes

[i] See: Fredrickson & Kahneman, 1993; Kahneman, 1999, 2000a, 2000b; Kahneman, Fredrickson, Schreiber & Redelmeier, 1993; Kahneman & Snell, 1990, 1992; Sarin & Wakker, 1997.

[ii] This section draws on a number of secondary sources, which are cited only when giving direct quotes: Georgescu-Roegen (1968, 1973); Niehans (1990); Stigler (1950a, 1950b). Most of what I say is drawn from two and sometimes three of these sources.

[iii] For an overview of this new approach, see Layard (2003).

[iv] In their 'normative analysis,' Kahneman, Wakker & Sarin (1997) assume only ordinal measurement (p.389). But these ordinal measures can be transformed into interval ones through the use of conventional scaling procedures (Kahneman, 2000a, 680-681).

[v] There are, perhaps, ultimate 'stop' signals. These include reflexes, which need not be mediated by central. If you put your hand on a burning stove your hand will 'go' before the signal gets anywhere near your brain. This stop point, however, occurs when a particularly kind of stimulation reaches the level of realized physical danger.

[vi] One issue, that will not be dealt with here, is that maximizing personal experienced utility is *not* the utilitarian principle, which concerns the social welfare. It remains logically possible, therefore, that decision utility could be a better approximation to the Benthamite principle.