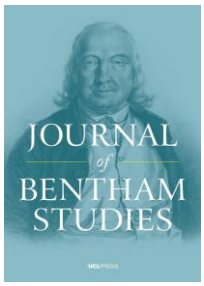

UCLPRESS

Journal of Bentham Studies

Jeremy Bentham: Ogre or Prophet?

Ken Binmore [1,*]

How to cite: Binmore, K. 'Jeremy Bentham: Ogre or Prophet?.' *Journal of Bentham Studies*, 2011, 13(1): 2, pp. 1–24. DOI: https://doi.org/10.14324/111.2045-757X.039.

Published: 01 January 2011

Peer Review:

This article has been peer reviewed through the journal's standard double blind peer review.

Copyright:

Open Access:

Journal of Bentham Studies is a peer-reviewed open access journal.

*Correspondence: k.binmore@ucl.ac.uk
[1] UCL, UK

UCL Bentham Project
Journal of Bentham Studies, vol. 13 (2011)
DOI: 10.14324/111.2045-757X.039

Jeremy Bentham: Ogre or Prophet?

KEN BINMORE

Economics Department, University College London

k.binmore@ucl.ac.uk

> Any man of sound head, and practised in wielding logic with a scholastic adroitness, might take up the whole academy of modern economists, and throttle them between heaven and earth with his finger and thumb, or bray their fungus heads to powder with a lady's fan.[1]

In his novel *Hard Times*, Charles Dickens encapsulates one view of Jeremy Bentham by representing him in the person of Mr Gradgrind, who tells little Louisa never to wonder: 'By means of addition, subtraction, multiplication, and division, settle everything somehow, and never wonder'.

This kind of bad press continues to this day in works like A. N. Wilson's *Victorians*, in which the author applies the word *Benthamite* to any economic development of which he disapproves, so that Bentham is simultaneously held responsible not only for all the crimes of authoritarian socialism, but also for the excesses of laissez-faire capitalism.[2]

I see no hope of countering this kind of mindless prejudice. As a supposedly heartless technocrat myself, I know that anyone intolerant of logical contradiction and willing to use mathematics in reasoning about human affairs still gets the same treatment today from do-gooders who resent the impracticalities of their utopian schemes being exposed by rational analysis. Unable to reply in kind, they resort to misrepresentation and character assassination. Bentham's deliberately provocative style doubtless fanned the flames of the criticism to which he was subjected in his own time, but the disgraceful treatment meted out to Edward Wilson in recent years for expressing his sociobiological views shows that the soft answer sufficeth not to turn away the wrath of utopians confronted with scientific facts that they dislike.

[1] Thomas de Quincey, *Confessions of an English Opium Eater*, London, 1821.
[2] A. N. Wilson, *The Victorians*, New York, 2003.

I am more interested in the fact that Bentham's reputation among modern scholars with no irrational axe to grind should be so low. His name is always twinned with that of John Stuart Mill when credit for utilitarianism is given, with the unspoken implication that the eccentric Bentham may have hit upon the phrase 'the greatest good for the greatest number', but Mill was the serious thinker who provided utilitarianism with sound intellectual foundations.[3]

It is true that Bentham was eccentric. The continuing display of his mummified corpse in the foyer of University College London as specified in his will is the most obvious of his many oddities. His obsession with building a circular prison in which a warder placed at the center could simultaneously monitor the behaviour of many prisoners is another. But it makes no sense to judge a creative genius by what we perceive as their follies. Otherwise it would be necessary to condemn Isaac Newton for his attempts to advance alchemy and numerology. As a famous old mathematician once explained to me after we had listened to some criticism of the work of a recently dead colleague: a man's achievements should be judged by his f_+, which is what is left after throwing the negative values of a function away.

John Stuart Mill

I am sure that John Stuart Mill had no evil intentions in writing an assessment of Jeremy Bentham after his death, in which he drew a parallel between Bentham, and the poet and would-be philosopher, Samuel Coleridge.[4] However, this assessment set in stone what still remains the modern view of Bentham: a minor celebrity in his time, but not a philosopher of the first rank.

Mill was particularly exercised by Bentham's refusal to be influenced by the school of German idealists, who were philosophically fashionable in England at that time. But it seems to me that, in such passages, Mill criticises Bentham for not being subject to his own weaknesses. After all, how can one follow up the insights of the Scottish Enlightenment as represented by David Hume, and simultaneously give

[3] I am grateful to my colleague Fred Rosen for pointing out that Bentham probably derived this formula neither from Hutcheson nor from Leibnitz, but from Beccaria's *Crimes and Punishments*. Bentham himself attributed his utilitarian ideas to Helvetius. See R. Shackleton, '"The Greatest Happiness of the Greatest Number: The History of Bentham's Phrase"', *Studies on Voltaire*, vol. 90 (1972), pp. 1461-1482.

[4] 'Bentham' and 'Coleridge' in *The Collected Works of John Stuart Mill, Vol. X: Essays on Ethics, Religion and Society*, ed. J. M. Robson, F. E. L. Priestley, and D.P. Dryer, Toronto, 1985, pp. 75-116, and 117-164. Henceforth *CWJSM*.

credence to the categorical denials of this approach embraced by Immanuel Kant? Such fudging was totally alien to Bentham, who did all his own thinking for himself and followed the logic wherever it led, however unpopular or unfashionable his conclusions might prove to be.

Nor does the view that Mill came up with intellectual foundations for utilitarianism that Bentham was unable to provide survive serious scrutiny. The much quoted chapter in which Mill supposedly provides such a foundation actually reduces to an ineffectual attempt to prove that what people want is happiness.[5] But what we need to know is what happiness is. How is it measured? Why should we add one person's happiness to another's? And so on. Nor is his attempt to show that the utilitarian creed is necessarily libertarian any more successful.[6] In both cases, Mill allows his conviction of the truth of the conclusion he is determined to reach to overcome his critical faculties in a manner that would have been impossible for Bentham.

This isn't to say that Bentham didn't also subscribe to convictions for which he was unable to provide a rigorous defence. But when he finds himself in this position, he doesn't deceive himself or others by offering the kind of waffle that he so condemned when he found it in the works of Blackstone and others—he simply explains that the proposition in question is to be treated as axiomatic. On what he calls the Principle of Utility, for example, Bentham observes that a proof is as 'impossible as it is needless'.[7]

David Hume.

My own view is that the time has come for the history of thought to accord Jeremy Bentham a more prominent position in the line of scientific philosophers that starts with Aristotle, and was continued by the likes of Epicurus, Hobbes, and Hume.

I think the link with the great David Hume is particularly important, since both Hume and Bentham were sufficiently ahead of their time that it is only with the advent of modern game theory that lesser minds have been able to find a framework within which their insights can sit comfortably.

[5] 'Utilitarianism' in *Ibid*., pp. 203-60.
[6] 'On Liberty' in *CWJSM Vol. XVIII: Essays on Politics and Society, Part I*, Toronto, 1977, pp. 213-310.
[7] *An Introduction to the Priniciples of Morals and Legislation*, ed. J. H. Burns and H. L. A. Hart, Oxford, 1996 (*The Collected Works of Jeremy Bentham*), p. 13. Henceforth *IPML* (*CW*).

My intention in this paper is to pursue this point by discussing Bentham's insights in the light of the work of modern economists like John Harsanyi, who have used game theory to take up where Hume and Bentham left off. In the process, I hope it will become clear that the sharp distinction between egalitarianism and utilitarianism perceived by modern philosophers is not something that can be traced back to Bentham, who has a good claim to have fathered both ways of looking at the world.

Three Questions

There are three questions that no utilitarian can evade:

- What constitutes utility?
- Why should individual utilities be added?
- Why should I maximize the sum of utilities rather than my own?

The early utilitarians had little to offer in answer to the first and second of these questions. With characteristic frankness, Bentham says, '[t]hat which is used to prove everything, cannot itself be proved'.[8] As for the additivity of happiness, this is quaintly described as a 'fictitious postulatum'.

Mill sometimes endorses this position, but also offers a halfhearted attempt at providing a proof of utilitarianism, which consists of a chapter devoted to the claim that what people desire is happiness.[9] Having established this proposition to his own satisfaction, he then rests on his laurels—apparently not feeling the need to tackle the second question.[10]

Sidgwick seems uninterested in foundational questions, but agrees with Bentham that the good is an 'unanalyzable notion'.[11] Only Edgeworth[12] is an exception to this Victorian intellectual vacuum, anticipating with his insurance

[8] *IPML*, p. 13
[9] *CWJSM*, pp. 209-26.
[10] All he offers on the second question is the observation: 'Each person's happiness is a good to that person, and the general happiness is therefore a good to the aggregate of all persons'.
[11] H. Sidgwick, *The Methods of Ethics*, Indianoplis, 1907.
[12] F. Edgeworth, *Mathematical Physics*, London, 1881.

argument the defense of utilitarianism offered nearly a century later by William Vickrey and John Harsanyi.[13]

The third question is more interesting from the historical point of view, because I believe that Bentham's position on this subject differs radically from that of most of his utilitarian successors. His focus on the practicalities of the law leads him to a position very close to that of David Hume on how social conventions are maintained in human societies.

Skyhooks

Neither Hume nor Bentham are willing to have any truck with what have nowadays come to be called 'skyhooks' by evolutionary theorists.[14] These are metaphysical or supernatural entities that utopians conjure from nowhere to hold aloft their castles in the air. Bentham is particularly scathing on this subject. Everybody knows, for example, of his rejection of the notion of imprescriptible natural rights as 'nonsense upon stilts' in his brilliant commentary on the French Declaration of the Rights of Man, but he is equally stout in denying all notions of natural law, however attractive the packaging in which it may be wrapped.[15] As he observes 'Instead of the phrase, Natural Law, you have sometimes, Law of Reason, Right Reason, Natural Justice, Natural Equity, Good Order'.[16] Only the names of such skyhooks have changed in modern times. For example, modern Kantians like Harsanyi or Rawls appeal to Moral Commitment or Natural Duty to explain why people should follow their prescriptions.[17] But for Hume or Bentham, one might as well appeal to Mumbo-Jumbo, or one of the many other supernatural beings that mankind has invented over the years.

Punishment

Bentham was not nearly so subtle in his assessment of how morality really works in actual societies as David Hume, but his insistence that the fear of punishment is at the root of moral behaviour puts him very much in the modern camp of those who see the

[13] J. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge, 1977.
[14] Daniel Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, London, 1995.
[15] *Rights, Representation and Reform: Nonsense Upon Stilts and Other Writings on the French Revolution*, ed. P. Schofield, C. Pease-Watkin and C. Blamires, Oxford, 2002 (*CW*), pp. 317-401.
[16] *IPML*, p. 27n.
[17] Harsanyi, *Rational Behavior*; J. Rawls, *A Theory of Justice*, Oxford, 1972.

folk theorem of repeated game theory as fundamental in explaining what it means to enjoy the rights and duties implicit in a social contract . As he observes 'Without the notion of punishment… no notion can we have of either *right* or *duty*.[18]

His later concerns about the abuse of power by those in government similarly echo some of David Hume's concerns. Everyone in society—including popes, kings, presidents, judges and policemen—needs to be properly incentified by the fear of punishment if they are to carry out their duties to the benefit of the society as a whole, rather than in pursuit of their own private ends.

I think this feature of Bentham's work has been badly neglected by posterity. His pragmatic view of human nature is seen as crude and demeaning when compared with the grandiose refinements proposed by philosophers who are more adept at telling us what we like to hear. But I think that the folk theorem provides a vindication of both Bentham and Hume on this subject. Sages from Confucius onwards have identified *reciprocity* as the secret of human sociality—and reciprocity works because those who fail to reciprocate are punished.[19] As Hume understood better than Bentham, the punishments are commonly much more subtle than the legal remedies on which Bentham concentrated, but Bentham was correct in identifying punishment as the crucial factor without which a moral code cannot be sustained.

The rest of this paper is structured by taking the three questions that open this section one by one, looking at what modern scholarship has made of them with a view to assessing Bentham's foresight. In undertaking this task, I was surprised to find that he was far less doctrinaire than I had always taken for granted—and much closer to my own position than I had thought possible.[20]

What is Utility?

In this section, I offer a brief sketch of the history of utility theory with a view to explaining the origins and tenets of the modern theory.

[18] *A Comment on the Commentaries and a Fragment of Government*, ed. J. H. Burns and H. L. A. Hart, London, 1977 (*CW*), p. 495n.
[19] What of reward? In modern thinking, failing to secure a reward counts as a punishment—an opportunity cost. Since Bentham invented the notion of an opportunity cost, it is disappointing to find numerous passages in his work in which he talks simultaneously of maximizing a benefit and minimizing a cost, although one can only optimize on one dimension at a time.
[20] In my book *Natural Justice*, New York, 2005—a title of which Bentham would not have approved—I offer a non-technical account of my own take on both utilitarianism and egalitarianism.

The word *utility* has always been difficult. Bentham himself 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.[21] The emergence of modern utility theory has only served to multiply the philosophical confusion. For example, Amartya Sen denied that John Harsanyi can properly be counted as a utilitarian at all, because Harsanyi interpreted utility in the modern sense of von Neumann and Morgenstern rather than in Bentham's original sense.[22]

But I see no reason why we should suppose that Bentham would have rejected the modern theory of utility if he could have had foreknowledge of its existence. It is certainly very much closer to his idea that we should root a moral theory in the real wants and aspirations of human individuals than the approach of those modern utilitarian philosophers who compile lists of criteria that supposedly determine what the good life ought to be. I fear that Bentham would have had little patience with paternalists who think they know better what is good for you and me than we do ourselves. He even coined a name for this kind of dogooder. He called them *ipsedixists*—those who think their own aspirations for society are somehow automatically superior to the aspirations of others.

Pleasure or pain?

Bentham perhaps thought that some kind of metering device might eventually be wired into a brain to measure how much pleasure or pain a person was experiencing. This is a view that economists in the early part of the twentieth century learned to lampoon mercilessly. It is true that such a naive theory of human motivation creates many difficulties, but it does not seem to me to deserve the derision it was accorded. After all, we all now know of 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. However, once the so-called 'marginalist revolution' had taught economists that their favourite theorems did not need the

[21] *IPML*, p. 11n
[22] A. Sen, 'Welfare Inequalities and Rawlsian Axiomatics', *Theory and Decision*, vol. 7 (1976), pp. 243-262. I suppose it is hopeless 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*. Following Ayer, 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. See A. J. Ayer, 'The Principle of Utility', *Jeremy Bentham and the Law: A Symposium*, ed. G. W. Keeton and G. Schwarzenberger, London, 1948, pp. 245-259.

cardinal utility functions with which they were traditionally proved, it became fashionable to denounce such utility functions as meaningless inventions.[23]

I suspect that one reason for this sea change was that it then became possible to assert that the interpersonal comparisons of utility that are necessary for utilitarianism to make sense could then also be denounced as meaningless. Even today, Lionel Robbins is still quoted as an authority for such fallacious claims, although von Neumann had already created what is now regarded as an entirely sound theory of cardinal utility at the time he was writing.[24]

Revealed preference

Critics of modern utility theory usually imagine that economists still hold fast to the primitive beliefs about the way our minds work that are implicit in the work of Bentham and Mill, but 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 behaviour.

This does not mean that economists believe that our thought processes have nothing to do with our behaviour. They know perfectly well that human beings are motivated by all kinds of considerations. People care about pleasure, and they care about pain. Some are greedy for money.[25] Others just want to stay out of jail. There are even saintly people who would sell the shirt off their back rather than see a baby cry. Economists accept that people are infinitely various but accommodate their infinite variety within a single theory by denying themselves the luxury of speculating about what is going on inside their heads. Instead, they pay attention only to what they see people doing.

The modern theory of utility therefore abandons any attempt to explain *why* people behave as they do. Instead of an explanatory theory, economists rest content with a descriptive theory, which can do no more than say that a person will be acting

[23] 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.
[24] L. Robbins, 'Inter-Personal Comparisons of Utility', *Economic Journal*, vol. 48 (1938), pp. 635-641; J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton, 1944.
[25] Bentham suggests at one point that we should measure felicity in money, as in modern cost-benefit analyses, but who would want to argue that an extra dollar is worth the same to a billionaire as to a beggar?

inconsistently if he or she did such-and-such in the past, but now plans to do so-and-so in the future.

Such a theory is rooted in observed behaviour. Following Samuelson, it is therefore called a theory of 'revealed preference', because the data it uses in determining what people want is not what they say they want—or what paternalists say they ought to want—but observations of what they actually choose when given the opportunity.[26]

Rationality as consistency

Oskar Morgenstern famously turned up at von Neumann's house one day in the early 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 irrelevant to morality, but it is hard to think of a more fundamental issue than who should bear what risk. Utilitarians who want to use the kind of insurance argument employed by Edgeworth or Harsanyi in defence of their position certainly have no choice but to accept that attitudes to risk are basic to their approach. Bentham's insistence on prioritizing security fits neatly into the same package.[27] 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.

Philosophers sometimes claim that rationality should mean more than mere consistency, so that some utility functions can be dismissed as being less rational than

[26] P. Samuelson, 'A Note on the Pure Theory of Consumers' Behaviour', *Economica*, vol. 5 (1938), pp. 61-71.

[27] Although taking the modern view requires transferring security questions from an analysis of what is optimal to an analysis of what is feasible. That is to say, security questions are built into each citizen's utility function, and therefore do not need to be considered separately when maximizing the sum of utility.

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, it would not be *irrational* for him to prefer the destruction of the entire universe to scratching his finger, because rationality is about means rather than ends.

Determining your utility

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

Pick two outcomes, *WIN* and *LOSE*, that are respectively better and worse than than any outcome that we need to discuss. (One can think of *WIN* and *LOSE* 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 *LOSE*, and 100 utils to *WIN*.

Suppose we now want to find David Hume's utility for scratching his finger. For this purpose, consider a bunch of (free) lottery tickets in which the prizes are either *WIN* or *LOSE*. As we offer Hume lottery tickets with higher and higher probabilities of getting *WIN* as an alternative to scratching his finger, he will eventually switch from saying no to saying yes. If the probability of *WIN* on the lottery ticket that makes him switch is 75%, then von Neumann and Morgenstern's theory says that scratching his thumb should count as being worth 75 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 *LOSE*, and 212 utils to *WIN*. 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 finger worth 75 utils on the old scale is worth 167 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 in predicting the behaviour of laboratory subjects as any comparable alternative proposed by the new school of behavioural economists, it cannot be said that it predicts very well in absolute terms.

Why Add Utilities?

To add the utilities of the citizens of a society is to take for granted that utility can be compared across individuals. It was this possibility that economic satirists like Lionel Robbins were most anxious to deny when heaping ridicule on the idea that cardinal utility scales could be meaningful.

Modern students of economics are still sometimes taught the dogma that interpersonal comparison is impossible, although von Neumann and Morgenstern's theory of cardinal utility is now fully accepted. The dogma is able to survive because the von Neumann and Morgenstern theory makes no assumptions about how people interact. Trying to use the raw utility scales the theory assigns to different people to assess their comparative welfare therefore makes no more sense than trying to tell which of two rooms is warmer without knowing whether the thermometers in the two rooms are both graduated in the same kind of degrees.

Bentham was making a similar point when he observed that adding the utilities of different individuals is like adding so many apples to so many pears.[28]

Ideal observer?

One objection to the idea that interpersonal comparison of utility can be possible is the claim that it is inconsistent with the principle of revealed preference. One is asked how such comparisons could be revealed by the choice behaviour of individuals.

John Harsanyi's cited answer is simple.[29] We reveal how we make such comparisons whenever we use a fairness criterion to decide who should get how much of some surplus. His own favourite example is of someone who is unable to use an expensive opera ticket, and so must now decide which of two friends would most enjoy it as a gift.

[28] ‘"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’. See J. R. Dinwiddy, *Bentham: Selected Writings of John Dinwiddy*, ed. W. Twining, Stanford, 2004, p. 49, and M. Mack, *Jeremy Bentham: an Odyssey of Ideas*, New York, 1963, p. 244.
[29] Harsanyi, *Rational Behavior*.

Harsanyi pursues this idea by imagining an 'ideal observer' or an 'impartial spectator' who makes such judgments on behalf of a society.[30] The decisions of the ideal observer are assumed to satisfy the consistency requirements of the von Neumann and Morgenstern theory. But Harsanyi confounds the traditional pessimism of the economics profession on this front by pointing out something that ought to have been obvious; one can add extra assumptions to von Neumann and Morgenstern's requirements.

Harsanyi's extra assumptions amount to requiring that the ideal observer be impersonal, in the sense that he takes account of nothing but the preferences of the citizens of his society when making decisions. The additivity of utility—Bentham's fictitious postulatum—is then an almost trivial consequence.[31]

When making decisions about who should bear what risks, the fact that the von Neumann and Morgenstern's assumptions require that the ideal observer maximize his *average* utility forces him to maximize a weighted sum of the utilities of the citizens of his society. The weights register how the ideal observer feels it appropriate to rescale each citizen's individual utility scale to make them comparable. After this rescaling, the ideal observer then acts as a utilitarian by adding together each citizen's rescaled utility to judge the worth of any reform.

The insurance argument behind this conclusion is more apparent in Harsanyi's second defence of utilitarianism.[32] In this second defence, he abandons the metaphor of an ideal observer in favour of his own version of the Rawlsian original position, which he invented independently of Rawls.[33]

The original position.

The original position is a hypothetical standpoint from which to evaluate the fairness of different ways of organizing a society.

All citizens imagine themselves behind a veil of ignorance that conceals their role in society. They then ask on what bargain they would agree in this imagined state of ignorance. Since their ignorance reduces everyone to a state of equality, they will

[30] *Ibid*., p. 49.
[31] Ken Binmore, *Just Playing: Game Theory and the Social Contract vol. II*, Cambridge, 1998, Appendix B.
[32] Harsanyi, *Rational Behaviour.*
[33] Rawls, *Theory*.

all agree on what kind of society is optimal. Harsanyi points out that this society will necessarily be utilitarian, because each citizen will wish to maximize his or her average utility on the assumption that he or she is equally likely to end up occupying any of the possible roles in the society on which agreement is reached in the original position.

Rawls famously derives an egalitarian conclusion from the same hypotheses, but his analysis does not survive serious scrutiny, since it depends on rejecting the \VNM\ theory in favour of the use of the maximin principle, which makes sense only in two-person, zero-sum games.[34]

Interpersonal comparison?

Harsanyi therefore offers two related arguments that offer an explanation of why utilities should be added, but neither argument solves the more basic problem of how and why utilities can be compared across individuals.[35]

In his first argument, the standard of interpersonal comparison is taken to be that of some mythical ideal observer, but how are we poor mortals to guess what standard of interpersonal comparison such an ideal observer would nurse in his bosom?

His second argument seems more promising on this front, but it turns out that Harsanyi's veil of ignorance is to be taken to be so thick that citizens in the original position forget even the standards of interpersonal comparison that operate in their current society. In this Kantian limbo, each citizen must construct a new standard of interpersonal comparison. Harsanyi then appeals to a dubious rationality principle—the so-called Harsanyi doctrine—which asserts that rational people in exactly the same situation will necessarily think exactly the same thoughts. In particular, they will subscribe to the same standard of interpersonal comparison.

But even if the Harsanyi doctrine were sound, we poor mortals would be no better off than with Harsanyi's first argument, since we have no more idea of what standard of interpersonal comparison an ideally rational person would construct in Harsanyi's original position than we do of the standard of interpersonal comparison to be attributed to his ideal observer.

[34] *Ibid*.
[35] Harsanyi, *Rational Behavior*.

Although commentators commonly overlook this point, both of Harsanyi's arguments therefore fail to solve the problem of how and why a consensus is to be established on how utilities can be compared across individuals.[36]

Naturalizing the argument

My guess is that Bentham would have been exasperated by the metaphysical aspects of both of Harsanyi's arguments. He would have denounced the ideal observer as more nonsense on stilts. The original position would similarly have been dismissed as an idle fancy.

Even if citizens were willing to go through the intellectual acrobatics required by Harsanyi's formulation of the original position, why should anyone feel bound to honor the hypothetical deal that hypothetically would be reached if the citizens of a society were to bargain in an hypothetical state of ignorance? Harsanyi's answer is that the citizens have an unexplained 'moral commitment' to honor the deal. Rawls says that they have a 'natural duty' to do the same. But we have already seen what Bentham thought of such skyhooks.

However, perhaps Bentham would have looked more favourably on my own naturalistic reinterpretation of Harsanyi's second argument. Hume thought it impossible to prove the principle of scientific induction, but that we are stuck with proceeding as though it were true because this is the way our brains work. I think the same about the original position.

Bentham would doubtless have argued that it is as impossible to find a rational justification for the original position as it is to prove the principle of utility, and I agree. But I would add the Humean proviso that we have no choice but to employ something similar to the original position, because the original position embodies the deep structure of the fairness norms with which evolution has equipped us for the purpose of solving the myriads of everyday coordination problems of which social life largely consists.[37]

The immediate point is that such a naturalistic reinterpretation of the original position allows us to dispense altogether with the skyhooks that Bentham so despised.

[36] Harsanyi unwittingly facilitates this mistake by renormalizing everybody's utility function at an early stage so that each of any citizen's utils is worth the same as any other citizen's utils. But such a renormalization would not be possible if the original utility functions were not comparable in the first place. See Harsanyi, *Rational Behavior*.

[37] Ken Binmore, *Playing Fair: Game Theory and the Social Contract, vol. I*, Cambridge, 1994; Binmore, *Just Playing*.

In particular, we no longer need to think in terms of some ideally rational standard of interpersonal comparison. When people use the device of the original position in everyday life to solve a coordination problem, they use the standards of interpersonal comparison that have evolved in their culture for this purpose.

I guess that Bentham would have looked askance at both the moral relativism implicit in this standpoint and its acceptance that history matters, but he surely would have welcomed the implication that determining the relevant standard of interpersonal comparison when applying his principle of utility is a matter of empirical observation.

Equity

The fact that Rawls[38] claimed that people would agree to an egalitarian social contract in the original position has been mentioned already. Since Rawls wrote his celebrated *Theory of Justice* explicitly to provide a reasoned alternative to utilitarianism, it may seem perverse for me to suggest that a modern Bentham might have found himself supporting Rawls rather then Harsanyi in the rather bad-tempered little debate that followed their discovery of each other's work.

I offer this suggestion, because Bentham repeatedly argues that the principle of utility will tend to generate equitable distributions of economic surpluses.[39] The following quote is typical: '[t]he less unequal the distribution of the external instruments of felicity is---the greater, so security be unshaken, will be the sum of felicity itself'.[40]

This conclusion certainly holds when money is to be split between two identical individuals, provided that their marginal utility for money decreases in the manner Bentham was perhaps the first to describe in the following aside: '[t]he quantity of happiness produced by a particle of wealth will be less and less at every particle'.[41]

But he clearly believed maximizing total utility will generate equal outcomes over a much larger domain than moden analysts would accept. He did not overlook altogether the fact that his principle of utility will often call for the sacrifice of the few for the sake of the many, but one has to search his writings with a toothcomb for a

[38] Rawls, *Theory*.
[39] Bowring, ii. pp. 267-74.
[40] Bowring, ii. p. 272.
[41] The utilitarian sum $u(x)+u(1-x)$ is largest when $u'(x)=u'(1-x)$. If $u'(x)$ is a strictly decreasing function of x (decreasing marginal utility), then $x=1-x$, and so $x=\half$. Bowring, iii. p. 229

suitable reference.[42] But what if he had appreciated the extent to which utilitarianism implies unequal outcomes? Would he have stood by the 'addibility of happiness' or would he have considered the use of a Rawlsian social welfare function instead?

Why be a utilitarian?

Sen and Williams comment adversely on the common failure of utilitarians to come clean on whether their theories relate to personal morality or public policy.[43] To a game theorist, this reduces to a question of enforcement. Do people maximize the sum of everybody's utility instead of their own because this is what they think they 'ought' to do, or because some powerful agency will punish them if they do not?

Harsanyi took the former view.[44] But if one follows Hume in believing that all 'categorical oughts' are just skyhooks without genuine binding power, Harsanyi's wordplay with the concept of moral commitment lacks all conviction. It is particularly hard to imagine Bentham swallowing such a notion. All his works, with their strong focus on legislation, take for granted that utilitarianism is about public policy. However, I think that critics seldom appreciate the sophistication of his views on this subject.

Mechanism design?

When quoting Bentham and Mill as authorities, welfare economists commonly take for granted that both followed the currently orthodox line that models government as an agency external to society.

A government's laws and tax policy, along with the conventions and common understandings inherited from the historical past, create the rules of a game for its citizens to play. Given that the government is able to enforce these rules, the citizens respond to the incentives and restrictions built into the rules by choosing strategies

[42] Egalitarians who are hostile to utilitarianism never tire of telling such stories. My favourite is the hypothetical case of a missionary who must be surrendered to cannibals so that his fellow missionaries can escape. For Bentham on the sacrifice of the few for the sake of the many, see Bowring, xi. p. 84: 'On every occasion in which the nature of the case renders the provision of an equal quantity of happiness for every one of them impossible, by its being a matter of necessity, he may sacrifice a portion of the happiness of the few, to the greater happiness of the rest'.
[43] *Utilitarianism and Beyond*, ed. A. Sen and B. Williams, Cambridge, 1982, p. 2
[44] Harsanyi, *Rational Behavior*. Although I do not see that this view is consistent with his emphatic defense of rule-utilitarianism as opposed to act-utilitarianism. It seems to me that that act-utilitarianism is what makes sense for the private morality option, and rule-utilitarianism for the public policy option. See Binmore, *Just Playing*.

that are in equilibrium—so that each player's strategy is an optimal reply to the strategies chosen by the other citizens.

Government officials evaluate this equilibrium using a social welfare function. Under ideal conditions, they design the rules of the civic game they create so that the equilibrium that is eventually played maximizes the value of their welfare function. If social welfare is measured as the sum of everybody's utility, then the government is said to be utilitarian.

Game theorists refer to this approach as *mechanism design* for reasons that Bentham would have appreciated: designing the rules of a civic game is no different in principle to designing a machine (or a prison).

Harsanyi's first argument can be coopted to defend such a public policy interpretation of utilitarianism. His ideal observer becomes an embodiment of an all-powerful, benign government—a philosopher-king in the original Platonic sense. If such a philosopher-king honors the rather mild assumptions that Harsanyi makes in his first argument, then he will necessarily be a utilitarian. However, I do not think that Bentham would have been at all satisfied with this welfarist conception of utilitarianism for the following reason.

Constitutional design?

The idea that a government is to be regarded as an incorruptible, enforcement agency that somehow exists outside society would have been as unacceptable to Bentham as to it was to Hume when he wrote:

> In constraining any system of government and fixing the several checks and controls of the constitution, every man ought to be supposed a knave and to have no other end in all his actions than private interest.[45]

In particular, government officials are no less players in the game of life than any other citizen. They respond to their incentives, just like ordinary citizens. We must expect that they will imperceptibly learn to put their own private interests before those of the public if given long enough to learn the ropes. Those who resist such corruption will gradually be supplanted by those who do not.

[45] D. Hume, *Essays and Treatises on Several Subjects: Essays, Moral, Political, and Literary*, i. Edinburgh, 1741, p. 49.

In his *Principles of a Constitutional Code*, and elsewhere, Bentham is explicit on this point. He tells us that the internal adversaries, against whose evil agency security is requisite, are the *unofficial* and the *official*. The latter are described as follows:

> The *official* are those evil-doers whose means of evil-doing are derived from the share they respectively possess in the aggregate powers of government. Among these, those of the highest grade, and in so far as supported by those of the highest, those of every inferior grade, are everywhere irresistible.[46]

In the same article, Bentham offers a very modern solution to this problem; government should be organized so that for each of its agents:

> the course prescribed by his particular *interest* shall on each occasion coincide, as completely as may be, with that prescribed by his *duty*.[47]

However, Bentham's practical suggestions on how this satisfactory state of affairs can be brought about seem naive to a modern reader, especially since Bentham was writing in the shadow of the great David Hume, who had already exposed the heart of the matter:

> When there offers, therefore, to our censure and examination, any plan of government, real or imaginary, where the power is distributed among several courts, and several orders of men, we should always consider the separate interest of each court, and each order; and, if we find that, by the skilful division of power, the interest must necessarily, in its operation, concur with the public, we may pronounce that government to be wise and happy.[48]

[46] Bowring, ii. p. 270.
[47] *Ibid*., p. 278.
[48] Hume, *Essays*, i. p. 50.

This, of course, is the recipe of 'checks and balances' on which the authors of the American Constitution relied, and of which Bentham was explicit in his approval.[49]

Social contracts as equilibria

I think it important for the history of thought to note that Jeremy Bentham followed David Hume in rejecting what we now call mechanism design as the appropriate pattern for constitutional reform, since it is still taken for granted that modern mechanism design is the appropriate model for such purposes in modern political economy—although nobody is able to explain how the constitution is to be protected from the abuses of the officials who supposedly enforce the rules of the game it delineates. The following quote from the constitution of the defunct Union of Soviet Socialist Republics would seem to say everything that needs to be said on this subject:

> Article 34: Citizens of the USSR are equal before the law, without distinction of origin, social or property status, race or nationality, sex, education, attitude to religion, type and nature of occupation, domicile, or other status.

Game theorists, mostly unknowingly, follow Hume on this front. He himself followed a long line of luminaries starting with Confucius in seeing reciprocity as the key to human sociality.

In modern game theory, Hume's insight is formalized by the folk theorem of repeated game theory, which shows that any outcome on which the citizens of a society might wish to contract, given the existence of adequate external enforcement, is also available as an *equilibrium* when the game being played is repeated indefinitely often, and the players care sufficiently about tomorrow to be incentified by the rewards or punishments that an action taken today may engender in the future.

This theorem makes it possible to see how human social contracts can survive without being suspended from imagined skyhooks. No external policeman is available to enforce the social contract of a sovereign state, but its social contract can nevertheless work, because its citizens police each other. This includes popes, kings,

[49] Although Bentham's praise of the 'Anglo-American United States' seems unlikely to find favour with modern Americans! See Bowring, ii. pp. 563-70.

judges, and members of the official constabulary, as well as ordinary citizens like the rest of us.

Game theory therefore offers a rigorous answer to the old question: who guards the guardians? The traditional answer offered by philosophers from Plato to Kant is that the chains of responsibility in an ideal state ascend upward to a philosopher-king at the top, who does his duty for reasons that somehow never get explained. Hume's alternative answer is that the chains of responsibility are closed. The checks and balances built into the constitution of an efficient state result in the guardians guarding each other.

An important consequence of this view is that it does not allow a constitution to be thought of as the rules of a game. In formal game theory, the players *cannot* break the rules of a game—the rules are taken to be inviolable. But as Bentham insists, insofar as the laws invented by human beings are honoured by the population at large, it is because those who choose to break them risk punishment.

If game theory is to be used in constitutional design, it is therefore necessary to abandon the methodology of mechanism design. Instead, the game that the citizens of a society are modeled as playing needs to be the immutable Game of Life, whose rules are determined by the laws of physics and biology, the facts of demography and geography, and everything else that it is beyond the power of man to alter. A social contract can then be identified with one of the equilibria of the Game of Life.[50]

Choosing a social contract—which I take to include a nation's constitution and civic code—then reduces to choosing an equilibrium from the infinite number of equilibria available in our complex Game of Life. This choice has historically been made by the impersonal forces of cultural evolution, but Bentham believed it possible for us to throw away the traditions built into our social contract, and to start all over again with a brand new social contract constructed according to rational principles.

Rights and duties.

Bentham believed, as I do, that the choice of a social contract *creates* the rights and duties that less rigorous authors commonly introduce as convenient skyhooks—

[50] Bentham's concern with security against the venality of officials is then absorbed into the question of whether a proposed social contract is indeed an equilibrium. As with individual attitudes to risk, security is therefore absorbed into the question of what is feasible.

usually with the prefix *natural* in the hope of evading the necessity of providing an explanation of their source.[51]

I hope I do not misrepresent Bentham's views by interpreting his notion of a duty as an action in the Game of Life that merits punishment of some kind if not carried out.[52] The duty to punish those who merit punishment is particularly important, since it is by this mechanism that the chains of responsibility are closed in an efficient social contract.

Bentham is also firm on what seems to me the insightful observation that there can be no rights without corresponding duties. In my own version of this proposition, I make the connection explicit by defining a right to be an action that you do not have a duty not to perform.[53]

I do not know to what extent others anticipated Bentham in this demystification of the idea of a right and a duty. My guess is that no predecessor could be so explicit, because they had no guiding principle sufficiently powerful that one might think of using it to design a social contract from scratch.

Political legitimacy.

To what extent is it possible to defend utilitarianism as a principle with which to solve the equilibrium selection problem that Bentham set himself? An obvious approach to this question is to seek to employ Harsanyi's version of the original position to this end. One can then appeal to the modern consensus that regards a government as legitimate if and only if it has a mandate from the people for the laws it enforces.

But what is a mandate? In practical terms, it means winning an election---often with the votes of only a smallish proportion of the full electorate, most of whom have only a vague idea of the policies proposed by the party for whom they are voting.

It seems to be generally accepted that winning such an election will not suffice in the case of constitutional issues, if only because of the importance of protecting minorities from potential oppression by a majority. That is to to say, constitutions need to be constructed in a fairer way than matters of everyday governmental policy. But what counts as fair?

[51] Bowring, ii. pp. 267-74.
[52] Such punishments need neither be judicial nor severe. The fear of social disapproval from one's peer group is usually an adequate incentive to prevent deviations from equilibrium play.
[53] Binmore, *Just Playing*.

One answer is that offered by Rawls and Harsanyi. A social contract is fair if it would be agreed in the original position. I endorse this judgment, but not for the Kantian reasons they offer. As commented earlier, I think that people commonly find the idea of the original position attractive because it captures the deep structure of the fairness norms that they regularly use in ordinary life when resolving everyday coordination problems–like who should give way to whom in a narrow corridor, or who should take how much of a dish in short supply.

So what social contract would be agreed in the original position? We have seen that Harsanyi's answer trumps Rawls' answer when we assume the existence of an external agency sufficiently powerful to enforce the agreed social contract. There is therefore no difficulty in arguing, for example, that a fair constitution for a trade union or a corporate entity will be utilitarian, because such sub-societies have the legal system of the society as a whole to serve as an enforcement agency.

But we cannot argue in the same way for the constitution of a sovereign state. So what happens if we analyze the bargaining problem faced by citizens in the original position when they know that any agreement they reach must be self-enforcing? Amongst other things, their agreement must then be proof against further appeals to the original position in the future.

This is not the place to review my analysis of this problem, although little more is involved than the observation that when nothing compels honouring an agreement to abide by the fall of a coin, then only agreements in which everybody is indifferent as to how the coin falls can be viable.[54] However, the conclusion provides a surprising vindication of Rawls'[55] moral intuition. Without external enforcement, the fair social contract agreed in the original position will be egalitarian, in the sense that each citizen's gain (measured in suitably rescaled utils) in moving from the current status quo will be equal.

Bentham an egalitarian!

Using modern ideas from game theory in trying to flesh out Bentham's ideas therefore allows us to flirt with a delightful paradox. Perhaps Bentham should have been an egalitarian rather than a utilitarian! I should hasten to explain that he would not have

[54] Binmore, *Just Playing*.
[55] Rawls, *Theory*.

been a deontological[56] egalitarian of the modern variety, but an egalitarian of the consequentialist school, who maximizes the *minimum* of all individual utilities rather than their *sum*.[57]

Bentham would not have needed to abandon his belief in the principle of utility to have embraced egalitarianism for this kind of reason. Citizens in the original position still evaluate their prospects behind the veil of ignorance by taking the average of the utilities of everybody in the society that they are creating. They therefore all act as good little utilitarians behind the veil of ignorance, but they see that the lack of external enforcement implies that simply choosing the social contract that maximizes average utility is not viable, because anyone who finds himself disadvantaged when emerging from behind the veil of ignorance will simply call for a renegotiation. They therefore agree instead on the best social contract that will survive such calls for renegotiation. This social contract is the efficient social contract in which everybody is treated equally according to the current standard of interpersonal comparison—which we have seen is a conclusion that Bentham was anxious to encourage.

I do not suppose that my suggestion that a modern Bentham might have embraced egalitarianism is likely to find much support, but perhaps the idea may provoke some reappraisal of the modern consensus on the real Bentham's beliefs and attitudes.

It is true that those who insist on rigorous reasoning from first principles have usually made their minds up in advance about the conclusions to which their analysis supposedly leads;[58] but I think that Bentham was one of a select few thinkers who are genuinely willing to follow the logic wherever it leads. And sometimes logic led Bentham to conclusions that modern utilitarians would find uncomfortably heterodox (if they did not choose to leave the relevant passages unread).

Ogre or Prophet?

[56] How did Bentham's invented term *deontology* come to have its present meaning as the doctrine that the Good must be explained in terms of the Right?

[57] The result would not be a society of clones for two reasons. The first is that utility would need to be interpreted as a gain over the current status quo, reflecting the fact that Bentham's supposed radicalism was decidedly muted. The second reason is that the feasible set would only contain equilibria of the Game of Life, reflecting Bentham's concern with stability and security.

[58] John Stuart Mill on the subject of liberty and utilitarianism is a case that readily comes to mind in the current context.

I guess it has to be admitted that Bentham was an ogre to the extent that he seems to have had no gift for personal empathy. He was certainly held in great affection by some of his young acolytes, but young men are always attracted by authority figures who shake the foundations of the Temple of the Philistines.

However, the real question for the history of thought is not whether Bentham's personality was deficient, or his behaviour eccentric, or his writing style quaint. All of these things are true, and more. But what really matters is whether he directed future research along profitable channels—and here I think posterity has failed to appreciate his prophetic insights adequately.

John Stuart Mill was a thinker worthy of respect, but Jeremy Bentham was a genuine landmark on the road to a scientific theory of social and political organization.