John R. LUCAS
Oxford University

Can the Theory of Games Save Mill's Utilitarianism?

John Stuart Mill's *Utilitarianism* engages our interest and sympathy because it is flawed. It reflects the crisis in Mill's life, when he lost his faith. He had been brought up by his father in the straitest tenets of utilitarianism, but had had nervous breakdown in early adult life from emotional ill-nourishment. Utilitarianism might work as a guide for the well-governing of India by James Mill and his colleagues, but gave little sustenance to the aspiring spirit of the Romantic Movement. It treats people as units, not individuals. It takes no account of the "projects" that people pursue, as Bernard Williams puts it, in his trenchant criticism of Utilitarianism.[1]

Each individual not only experiences pain and pleasure as things happen to him, but also is an initiator of action according to policies he has framed in the light of his view of himself and the world around him. Each of us uses the first person singular, not just to say "it is hurting me" or "I like that", but to frame intentions and carry them out in action. *Ego, ergo ago*. We are essentially agents, not just sentient beings – indeed, we could not be sentient unless we were typically able to respond to pleasure or pain in an appropriate fashion. Mill, the author of *On Liberty*, needed to register that each person was to himself the first person, and needed to respect that first-personal stance. But filial piety forbad his throwing over utilitarianism explicitly. He continues to be a utilitarian in his own eyes, maintaining that it has been misunderstood by its critics, and that as properly expounded by him, it is free from the defects its critics have seized upon. This makes for intellectual acrobatics of an intriguing kind, which others here will explore and exploit. What I want to do, however, is to offer a wider framework, based on the Theory of Games, which will put both halves of Mill's thought into context, where we can see how they, and their respective strengths and weaknesses, relate to each other.

The Theory of Games helps us understand our reasoning when we make decisions involving more than one person. It shows why I need to take account of other people's decision-making as well as my own, why what has happened in the past is relevant as well as what may happen in the future, and why my values need to develop to

---

[1]1. J.J.C. Smart & Bernard Williams, *Utilitarianism: for and against*; Cambridge University Press, Cambridge, 1973.

encompass our common good and not just my own individual good.

In the Theory of Games each decision-maker, or "player", has a number of choices, yielding a large number of "outcomes" according to the choices made by himself and other players. Thus if there are four players each with three possible courses of action, there will be 81 (i.e., 3x3x3x3) possible outcomes. Each outcome is evaluated by each player according to his system of values, and the value he assigns to it is called his "pay-off". The pay-off is normally expressed in numerical terms, with the suggestion that we are dealing with the cardinal, interpersonal utilities that utilitarians believe in, but there is no need to assume that they are always cardinal and interpersonal; for most purposes it is enough that each player can decide his order of priorities as between the various outcomes that may result from his and others' choices.[2]

In the games-theoretical framework, Act Utilitarianism is the limiting case of a one-person game, in which the utilitarian is the sole decision-maker, and decides so as to bring about that outcome which will have the highest pay-off. It has what Bernard Williams calls the "Government House" attitude; it is benevolent; it wants to do what will be best for its people; but it does not reckon that their acting according to their lights is something that should be accorded serious respect. It is benevolent, but it is benevolent despotism.

Once we allow that there are other people who make their own decisions, conflicts can arise, some of which yield puzzling results. The most famous and most familiar is the Prisoners' Dilemma. It was first discerned by Protagoras,[3] and greatly impressed Plato,[4] and later Hobbes, who made it the cornerstone of his argument for Leviathan. In its modern form it is due to A.W. Tucker.[5]

Game 1: Prisoners' Dilemma

In the matrix on the next page we represent two decision-makers, me on the left-hand side and the other chap on the top. We each have two choices, yielding four possible outcomes, each of which has

---

[2] This account is drawn from fuller ones in M. R. Griffiths & J. R. Lucas, *Ethical Economics*; Macmillan, Basingstoke, 1996, Appendix A, pp. 222-229; J.R. Lucas, *On Justice*, Clarendon Press, Oxford, 1980, ch.3, pp.35-71; J.R. Lucas, Responsibility, Oxford: Clarendon Press, 1993, ch.4. 4.7, pp. 69-72.

[3] Plato, *Protagoras*, 322ff.

[4] Plato, *Republic* II, 369-171.

[5] Tucker's formulation did not come out in a research paper, but in a classroom. As S. J. Hagenmayer wrote in The Philadelphia Inquirer ("Albert W. Tucker, 89, *Famed Mathematician*," Thursday, Feb. 2, 1995, p. B7) "In 1950, while addressing an audience of psychologists at Stanford University, where he was a visiting professor, Mr. Tucker created the Prisoners' Dilemma to illustrate the difficulty of analyzing" certain kinds of games. "Mr. Tucker's simple explanation has since given rise to a vast body of literature in subjects as diverse as philosophy, ethics, biology, sociology, political science, economics, and, of course, game theory."

its value – pay-off – for me (shown by the numeral at the bottom left of the outcome), and its pay-off for the other chap (shown by the numeral at the top right of the outcome).

	He wrongs me	He does not wrong me
I wrong him	1 Life for both of us is nasty, brutish and short, but at least I occasionally get some of his goodies. 1	0 Life for me is lovely. I enjoy the security of not being wronged by him with the liberty of wronging him whenever convenient Life for him is nasty, brutish and short, with no consolations whatever. 10
I do not wrong him	10 Life for him is lovely. He enjoys the security of not being wronged by me combined with the liberty of wronging me whenever he feels like it. Life for me is nasty, brutish and short, and even when I get the opportunity of taking advantage of him, I don't take it. 0	6 Life for us both is tolerable, but circumscribed. We both enjoy security from each other's depredations, but both are frustrated in the full exercise of our own personal potential. Life is comfortable, bourgeois and long, but lacking in authenticity; and we both suffer from *mauvaise foi*. 6

What the Prisoners' Dilemma establishes is the irrationality of "me-firstism". If we stick with the first person singular, we shall adopt a policy that in some situations will yield worse outcomes for each of us than we should obtain if we moved from the singular to the plural, and considered what was best for us all. We do better if we cooperate with one another, than if I, and everybody else likewise, thinks only of himself.

Game 2: Rule of the Road

M. Chevalier \ Mr. Knight	goes right	goes left
à droit	5 each passes other safely 5	0 collision 0
à gauche	0 collision 0	5 each passes other safely 5

The Rule of the Road shows the importance of conventions "Coordination Norms" in enabling players in a many-person game to concert their decisions so as to secure outcomes that they all prefer. In driving, in communicating, in dancing and in many other social activities, we need to coordinate our actions with one another, so as to concert our efforts and avoid collisions. Schematically we represent two motorists, Mr. Knight and M. Chevalier approaching each other, and needing to move over in order not to run into each other, by the matrix (with Mr. Knight's pay-offs in top right of each outcome, and M. Chevalier's in bottom left). Provided both go right, or both go left, they will pass each other safely: what is essential is that they do not each decide what he, on his own, thinks best, but both abide by some convention, or rule, or law, or mutual agreement. That is to say, I should not attempt to do whatever seems to me to be productive of the best consequences, but should reliably act in the way that other people expect me to act. I should drive on the left and not cut corners, give way when the other driver has the right of way, and press forward when I have, so that other drivers know where they are with me, and can plan their own movements accordingly. There is a necessary imperfection of information about the future actions of free agents in the absence of publicly avowed rules: norm-observance (deontology) is the key to coordination. A simple maximizing strategy is impossible, and each player must keep in step with others, usually by means of their all abiding by some relevant convention. Whatever the apparent attractions of consequentialism for the single operator, they are shown to be illusory, even by consequentialist standards, once the agent sees himself to be not a solipsistic loner, but one person among many, each needing to recognise others as initiators of action with minds of their own whose decisions can be anticipated only if they adhere to well-known rules.

Game 3: Battle of the Sexes

He \ She	goes to Alps	goes to sea
goes to Alps	8 “lovely for him; good for her” 10	4 “wish you were here too” 4
goes to sea	0 beastly for him; beastly for her 0	10 Good for him; lovely for her 8

In the Battle of the Sexes He and She want to spend their holiday together, but He would prefer to go mountaineering in the Alps, whereas She would rather they both spent it sunbathing by the sea. Since for either of them the second best is so much better than the third or fourth alternatives, it would pay either to settle for that if the very best appeared unattainable. And therefore it would pay the other to make it seem so. If She can throw a fit of hysterics and say she cannot abide the Alps and will not go there at any price, then He, if he is reasonable, will abandon his hopes of an Alpine holiday, and settle for the sea, which he would like twice as much as solitary mountaineering. But equally He may see that the moment has come to take a firm masculine line, and let the little woman face up to the realities of the situation, and either come along with him or go her separate way. And if once it becomes clear that this is the choice, She will have no option but to cave in, and buy a knapsack instead of a new bikini. It is thus irrational to be guided only by the pay-offs of the outcomes that are available at any one time, because that enables the other to manipulate one's choices. If I am to retain my autonomy, I cannot be altogether a direct consequentialist. Once you know that I am guided by consequences alone, you can induce me to do whatever you want by rigging the situation in such a way that by the time I come to make a decision the least bad outcome available to me is to fall in with your plans. Rationality requires, instead, that we extend our consideration over time as well as person.

It is often an advantage to be able to bind oneself absolutely, or equivalently to rule out certain options absolutely. The strategy of Mutually Assured Destruction only worked provided both sides believed that the other was not governed solely by consequentialist considerations, and really would retaliate if attacked, even though there would be then no advantage in doing so. In order to reinforce this expectation, mechanical devices were constructed which in the event of a nuclear attack would operate automatically without the possibility of being switched off by any consequentialist survivors. In

a less grisly way the whole logic of making and keeping promises is to ensure that some actions of an agent need not be altered simply by reason of factors, which had been future, becoming, by the effluxion of time, past. If we discount all past considerations we not only lay ourselves open to manipulation, but give only a partial account of the context in which our decisions are made, and from which they obtain their significance. I cannot be coherently oriented towards the future alone once I recognize that all my futures will one day be past.

Each of these arguments is a *reductio ad absurdum*. We start by assuming, as the classical economists did, that rationality can be defined in terms of maximising future pay-offs, and then show that even within its own terms, such a definition is self-contradictory. The Prisoners' Dilemma shows that he should take into account not only the existence but the interests and ideals of other people, and that it is irrational to ignore the collective point of view. The Rule of the Road shows that it is better to keep to the rules than to try, as the Act Utilitarians counsel, to perform the act that will have the best consequences: each of us should recognise that he is not the only pebble on the beach, that it is not for him to choose which course of events shall occur, and that often the best he can do is to fit in with what other people are likely to do. The Battle of the Sexes shows that it is irrational to have regard only to future outcomes; an agent has a past as well as a future, and should make up his mind what he is going to do with regard to what he has decided in the past as well as what will ensue in the future. Contrary to the static, solipsistic, future-oriented, exclusively individualistic standpoint of the classical economists, we are forced, by thinking about these three cases, to recognise that rationality is dynamic, leading us to take a longer temporal and wider personal view of what is involved in the decisions we are called on to take.

Mill would have welcomed the games-theoretical approach, although the arguments given here would not have provided him with all he needed – justice in particular cannot be secured by these arguments alone, but needs further arguments in which we address in the non-intimate second person (‘you’ rather than ‘thou’) those against whom adverse decisions are being taken.[6] Nevertheless, the games-theoretical approach would have reinstated the individual as an agent with his own projects and plan, and not just a unit capable only of experiencing pleasure and pain, and would also have offered a way of arguing from the premise that each desires his own happiness to the conclusion that we all ought to pursue the happiness of all.

Yet Mill might still have worried, as a pious Utilitarian, how he might persuade his father to conjugate. But the principle of conjugation was already accepted in that Act Utilitarianism considers future consequences and not only present ones. The pristine pleasure

[6] See more fully, J.R. Lucas, *On Justice*; Clarendon Press, Oxford, 1980, ch. 1; or “The Concept of Justice” at http://users.ox.ac.uk/~jrlucas/libeqsor/justice.html.

principle urges us to pursue present pleasure – *to paron hedu* – and it is only after much education and nagging that we move from simple hedonism to prudence. That move once made, it is incoherent to object on principle to considering yet other times and other persons.

The Theory of Games gives a better account of the individual decision-maker as being not simply an isolated *ego* with a truncated view of time, concerned only with the future, but as an agent with a past too, and as having reasons to identify himself with us, and with them and with you. But we should beware of over-conjugating, to the extent that we entirely lose the importance of the first-person singular now. Counsels of prudence can lead us to mortgage the present to the future, and far too often in the twentieth century the individual's interest was submerged in some imagined interest of a spurious collectivity. The twentieth century was made miserable by ideas of self-determination thought to be more important than concern for actual individuals and their security and freedom. "Better self-government than good government" was the motto. I remember as an undergraduate some fifty years ago arguing with Bernard Williams in Balliol JCR about Burma, which was in the process of being made independent, and liberated from the shackles of British Imperial rule. I was skeptical, and said that if I were a Burmese peasant, I should much prefer to be governed by a British civil servant, who might be distant and stand-offish, but would be impartial and fair-minded, and somewhat inclined to benevolence, than by the local *dacoit*, war-lord, who would speak the same language and have the same coloured skin, but would have no compunction in tyrannizing and oppressing me. Bernard would have none of it. He was in tune with the spirit of the age. But now, fifty years later, as information trickles out about the plight of the people under their present government,[7] my doubts are sadly vindicated. Mill, if he were living now, would have conjugated: but he would never have lost sight of the paramount importance of the first person singular in the present tense.

[7] See, for example, Pascal Khoo Thwe, From the Land of Green Ghosts: A Burmese Odyssey, London, (Harper Collins, 2002, or Flamingo, 2002, or Flamingo 2003, Harper Perennial, 2004).