"Summary of article by J.C. Harsanyi: Game and Decision-Theoretic Models in Ethics" in Frontier Issues in Economic Thought, Volume 3: Human Well-Being and Economic Goals. Island Press: Washington DC, 1997. pp. 113-116

Social Science Library: Frontier Thinking in Sustainable Development and Human Well-being

"Summary of article by J.C. Harsanyi: Game and Decision-Theoretic Models in Ethics"

Until the 1930s, a utilitarian ethical philosophy was widely accepted among economists. This philosophy often included the assumptions of cardinal utility and interpersonal comparability, i.e., that it was possible to measure an individual's utility, and to make quantitative comparisons of the utility experienced by different people. Since the "ordinalist revolution" of the 1930s, a majority of economists have rejected cardinal utility and interpersonal comparability, leading to considerable problems of reconstructing welfare economics in the absence of these foundational assumptions. A minority has argued for a return to a form of utilitarianism. In this article, John Harsanyi, the best-known of the "new utilitarians", argues that rational behavior implies the existence of cardinal utility functions for individuals, and a social welfare function for society. He also distinguishes his version of utilitarianism from other utilitarian and nonutilitarian philosophies.

SOCIAL UTILITY

Suppose that people respond rationally to situations like lotteries: that is, situations in which any of two or more outcomes can occur, with known (or subjectively estimated) probabilities. Literally buying a lottery ticket gives rise to an overwhelming probability of simply losing the price of the ticket, and a slight probability of winning a jackpot. Driving faster than the speed limit is also a lottery in abstract terms; it leads to some probability of arriving at the destination sooner, and increased probabilities of being stopped by the police or having an accident. "Rational" decision-making means that an individual is able to compare any two lotteries (either it is clear which one is preferred, or both are equally attractive); if the outcome of lottery A is at least as good as the outcome of B under every possible situation, then lottery A as a whole is at least as attractive as B; two lotteries that have the same prizes with the same probabilities are equally attractive; and if A is better than B, which is better than C, then some weighted average of A and C is exactly as good as B.

Any individual who is rational in this sense has an "expected utility" function, such that the expected utility of a lottery is the weighted average of the utility of the prizes, weighted by the probability of obtaining each prize. It is unique up to a linear transformation -- that is, once the zero point and unit of measurement have been chosen, the expected utility function is uniquely defined. This result was first proved by John von Neumann and Oskar Morgenstern in their pioneering work on game theory; the expected utility function is often referred to as the von Neumann-Morgenstern (vNM) utility function.

Building on this result, modern utilitarianism claims that all morality should be based on maximizing social utility, or a social welfare function, which is the sum, or average, of all individual utilities, when measured in the same units. To demonstrate this point, it is necessary to distinguish between an individual's personal preferences and moral preferences. Personal preferences are particularistic, giving more weight to oneself, relatives, and friends than to unknown other members of society; moral preferences are universalistic, giving the same weight to everybody's interests. Moral preferences exist independent of an individual's position in society; they would be equally applicable if an individual did not know who he or she was going to be, but had an equal probability of being in any social role.[1] Under these circumstances, the moral valuation of any situation can only be based on the unweighted average of its utility to every individual. Likewise, public policy, if made rationally, will maximize the policymaker's best estimate of (unweighted) average social utility.

Of course, calculation of average utility is not possible unless interpersonal comparisons of welfare can be made. Comparing the level of satisfaction of two individuals is not a trivial task, but neither is it meaningless. The statement, "he is less satisfied with his career than she is with hers" is difficult to evaluate unless we know them both well -- but, when referring to people we do know well, we frequently make and discuss such statements. It is easier to compare utilities if they are interpreted as measuring amounts of satisfaction, rather than preference orderings.

A common but mistaken objection to the use of vNM utility functions is that they merely express people's attitudes toward gambling, and thus have no moral significance. If we distinguish between the process utility (positive or negative) obtained from the act of gambling, and the outcome utility derived from the prizes (or losses), it is clear that the outcome utilities are what is important. Despite the definition in terms of lotteries, vNM utility functions depend only on outcome utilities: the description of rational decision-making, given above, implies that two lotteries differing immensely in process utility, but identical in outcomes, must be evaluated identically.

RULE UTILITARIANISM AND RAWLS' THEORY OF JUSTICE

It is important to distinguish two varieties of utilitarianism. Act utilitarianism asserts that the morally right action is the one that maximizes expected social utility in the existing situation, while rule utilitarianism requires a two-step process: first, define the moral rule that maximizes social utility in similar situations; second, act according to that rule. Since different moral rules are interdependent, rule utilitarianism requires adoption of a utility-maximizing moral code in general.

There are a number of drawbacks to act utilitarianism. It would require an impossible amount of calculation of utilities. It would deprive people of the incentives and assurances obtained from knowing that a given moral code was being followed. It would not allow the existence of any morally protected rights and obligations, nor any binding contracts and commitments, since such considerations could be overridden by a utilitarian calculation at any time. In sum, most of us would much prefer to live in a rule utilitarian world of stable moral codes -- which in itself is a utilitarian argument for rule utilitarianism!

Both varieties of utilitarianism are consequentialist ethical theories, defining morally right behavior ultimately in terms of its consequences for social utility. This provides a rational foundation for moral choices which is lacking in nonconsequentialist theories, such as John Rawls' theory of justice. Rawls attributes the principles of justice to a fictitious social contract, adopted under the "veil of ignorance", that is, without individuals knowing what role they will play in society. While this bears some resemblance to the view of moral value judgments presented above, Rawls then argues that a person operating behind the veil of ignorance would not maximize average social utility, but rather would choose to maximize the welfare of the worst-off members of society -- the maximin principle. This principle makes the value of any action or situation dependent on its worst possible outcome, not its expected value (which is a probability-weighted average of the value of all possible outcomes). In general, this is a poor guide to both practical and moral decisionmaking.

REASSESSING INDIVIDUAL UTILITIES

Several modifications and clarifications of individual preferences and utilities are required for the full development of a utilitarian ethics. Individual preferences based on mistaken or incomplete information do not correspond to a person's real interests; choosing to drink a glass of orange juice because you do not know that it contains poison does not mean that you prefer to be poisoned. Thus it is fully informed preferences that should be represented in utility functions. Likewise, malevolent preferences should be excluded; they cannot be rationally supported by a society based on benevolence toward individuals. In fact, all other-oriented preferences, even benevolent ones, should be excluded; failure to do so would mean that the welfare of the most popular individuals, with the largest numbers of well-wishers, would be counted disproportionately heavily in the social welfare function.

Benevolence toward another person does require us if possible to treat him as he wants to be treated. But it does not require us by any means to treat other people as he wants them to be treated. (In fact, benevolence toward these people requires us to treat them as they want to be treated, not as he wants them to be treated.) (704-705)

This implies that the social welfare function should be the sum, or unweighted average, of each individual's informed, self-directed preferences.

Notes

1. This argument, strongly reminiscent of Rawls' "veil of ignorance", was apparently developed, independently, three times in the 1940s and 1950s -- first by William Vickrey, second by Harsanyi, and finally by Rawls. (695, note 7)