When Ignorance Fails: a Critique of Harsanyi's Impartial Observer

Andrew Rennemo*

John Harsanyi tries to defend utilitarianism against its deontological critics by linking ethics, and utilitarian morality more specifically, with rationality. Harsanyi presents two arguments for utilitarianism that both lead to the same conclusion, namely, that a rational person will prefer a social system that maximizes average expected utility. However, this paper will argue that Harsanyi does not mount a convincing defense. To place Harsanyi's arguments in context, this paper will first define "social arrangement" and discuss the two dominant schools of thought, contractualist and consequentialist. It will introduce utilitarianism as a strain of the consequentialist school and outline its basic tenets. It will then present Harsanyi's argument for utilitarianism, and assess that defense as insufficient because of the tenuousness of several key assumptions.

A social arrangement is the set of institutions, rules, and regulations that serve as the framework of a society. Contractualist and consequentialist theories are the two schools of thought that try to evaluate such an arrangement. The contractualist tradition tends to focus more on an arrangements 'fairness,' paying comparatively less attention to its 'value.' It evaluates an arrangement by constructing a type of "potted history," a thought experiment to determine what ground rules for the creation of a society people would agree upon through negotiation with each other. Conversely, the consequentialist tradition conflates value and fairness in its evaluation of an arrangement. Welfarist theories, which measure the value of an arrangement by the expected welfare of all individuals living under it, are a prime example because they define social fairness as what is predicted to benefit everyone involved.

Utilitarian doctrine incorporates consequentialism, welfarism and impartiality. It is consequentialist because it places all moral weight on how the outcomes of societal rules and regulations affect the welfare of individuals living under them. It is welfarist because it argues a good and just society is one that maximizes the expected

welfare of all its members, either in average or in total. While utilitarianism has been associated with eternally evolving definitions of welfare, i.e. Bentham's hedonistic happiness, preference satisfaction, and then satisfaction of self-interested and well-informed preferences, the emphasis on its maximization remains constant. Further, utilitarianism is impartial because it assigns equal moral weight to the welfare of every person. As evidenced by Peter Singer's argument for giving to charity in "*Famine, Affluence, and Morality,*" a utilitarian actor must be as concerned for the well-being of a total stranger as he is for that of his own mother, brother, or spouse. The same goes for utilitarian social arrangements and component institutions. As R.E. Goodin notes, citizens are entitled to only one claim against them, namely, that they give equal concern in their utility calculations to the welfare of every person in society.

John Harsanyi attempts to defend (rule) utilitarianism against its many critics with two interwoven arguments, one axiomatic, the other contractualist, both of which link ethical behavior with rationality. The rational groundwork for this linkage is as follows. Harsanyi's first premise states that ethics is a theory of action that tells its adherents how they should behave and regard others, in trying to pursue the common interests of society. His second premise is that consistency must be characteristic of morality. Without consistency, there remains only intuition to guide ethical conduct, and people's intuitions are often determined by the highly arbitrary facts like place of birth and how they are raised. Since a normative discipline cannot be driven by arbitrary factors, an ethical theory by Harsanyi's definition must be a consistent one. Thirdly, rationality implies consistency as one of its main tenets. This is reflected methodologically in the fact that, as one of the primary axioms of rational choice theory, transitivity prohibits individual choice among paired alternatives from cycling illogically. Therefore, as Harsanyi states, utilitarianism is a doctrine worthy of defending because its modern version is the only ethical theory that determines morality through the use of rational tests and thereby gives morality the consistency it requires.

The first defense Harsanyi employs is an impartial thought experiment that follows in the social contract tradition and is largely indebted to the moral philosophy of Adam Smith. Harsanyi asserts every person contains within himself two separate preference rankings, one of self-interested preferences and the other of moral preferences. The relevant distinction here is that self-interested preferences are not widely other-regarding. Though Harsanyi acknowledges not every personal utility function will be wholly selfish, he notes such rankings will assign comparatively greater weight to the individual's own welfare and that of his loved ones and lesser weight to the welfare of strangers. However, moral preferences, represented

by a person's social welfare function, are a special, altruistic kind of preference for doing the "right" action, which Harsanyi says is the same as doing what benefits the common interest. Harsanyi states that different people will have different utility functions, but in theory will have the same social welfare functions. This is because, as Harsanyi argues with the presentation of his "similarity postulate" based upon Adam Smith's "mutual sympathy," in the absence of evidence to the contrary, we may assume the similarity of persons and attribute interpersonal differences to arbitrary factors like upbringing. These arbitrary factors aside, all people are made of the same "material," and will accordingly make objective interpersonal utility comparisons based on empathy with others and reach the same conclusion about what benefits society. In order to set the arbitrary interpersonal differences aside, the hypothetical individuals in the thought experiment must be motivated by their moral preferences. The way to act solely on moral preferences in reality is for an individual to try consciously to suppress self-interested inclinations and view a situation objectively. This may be modeled, Harsanyi argues, by having each hypothetical person in his thought experiment put himself in a position similar to Adam Smith's "impartial observer."

The thought experiment, originally introduced in Harsanyi's *"Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking"* and later elaborated upon in *"Morality and the Theory of Rational Behavior,"* proceeds as follows. Suppose a society consisting of n individuals, 1 to n, with each number corresponding to a social position in that society. An individual numbered "1" holds the highest social position, "2" the second highest, and so on. The levels of utility enjoyed by the people in social positions 1 through n are denoted as U1, U2, all the way down to Un. A given individual that wants to make judgments of value and fairness between two alternative social systems, such as between capitalism and state socialism, is called individual "i." To set aside morally irrelevant self-interested preferences and allow for a rational and objective decision, individual "i" assumes the position of the aforementioned "impartial observer." As such, he is ignorant of the social position and related interests he holds, and of those of his loved ones. The only thing "i" does know is what Harsanyi grants him in the "equiprobability postulate," namely, that he has the same probability, 1/n, of being each person in society, and thus of having each of the utility levels U1 to Un. This means that, when selecting between alternate social arrangements, "i" must act as if he is as likely to be a wealthy industrialist in a capitalist society as he is to be a poor shop clerk under state socialism.

Of the two schools of thought regarding how a rational person should make decisions under uncertainty, the maximin principle and the principle of ex-

pected utility maximization, Harsanyi argues that individual "i" should use the latter method. According to this approach, known as Bayesian decision theory, individual "i" calculates the expected utility of being in a given social position, "j," in society "x" by multiplying the expected utility attached to holding that position by the probability value, 1/n, that he will, in fact, hold the position. Due to the uncertainty, "i" selects the social system that will maximize the arithmetic mean of all individual utility levels in society, and thus the system that will maximize his own expected utility level. This social arrangement is represented in Harsanyi's Theorem T.

Harsanyi's second argument, his utilitarian representation theorem, is a technical representation of the first. It seeks to show formally that, if a person is minimally rational and minimally egalitarian, that person must have a utilitarian moral compass. As it is laid out in *"Morality and the Theory of Rational Behavior,"* this argument contains four axioms. The first is an individual rationality assumption, which states that a von Neumann-Morgenstern utility function may be used to represent the personal preferences of all n individuals in society. The second axiom assumes the "rationality of moral preferences," meaning at least one person in society, individual i, will be, in Harsanyi's words, as rational in determining the common interests of society as he is in looking after his own personal interests. Thirdly, there is a "Pareto optimality" axiom supposing that, to use Harsanyi's terminology, if there is at least one person that prefers a first alternative, A, to a second alternative, B, and no one has a contradictory preference, then individual "i" will morally prefer A over B. The final axiom is one of symmetry, requiring that the utility of all individuals be given equal weight. The conclusion Harsanyi derives from this axiomatic presentation is again Theorem T.

However, two key assumptions in Harsanyi's arguments are potentially faulty. Firstly, he equates a morally "right" action with what serves the common interests of society. Yet, this is often not the case. For when a pauper breaks a promise to repay a loan to a spiteful, selfish and wealthy lender, a strong theoretical argument could be made that he acted on a moral preference that society be mercifully just to the needy and was morally correct in doing so. This, however, is far from an action that best serves the common interest in the long term because it would set a damaging social precedent. The maxim that one should only keep promises that are easy to fulfill would fail the contradiction in conception test in Kant's four-step categorical imperative procedure because one could not live successfully in a society in which this was a universal law. Thus, given that morally "correct" action does not always coincide with what is best for society, it is unlikely that all people driven solely by their moral preferences in Harsanyi's thought experiment will converge on an altruistic

social system like one that maximizes average expected utility.

A second, very large assumption is that which underlies Harsanyi's equiprobability postulate. He assumes that the natural way to dissect a society is by its population. That is, when ignorant of the probability of being any individual in a society of n people, one should assign equal probability, 1/n, to each alternative. In fact, an equally, if not more, sensible approach would dissect a society by its social positions. There were, for example, unarguably more peasants than aristocrats in feudal France, just as there are more unskilled workers in a given capitalist country than wealthy capitalists. Acknowledging this fact means "i" accepting a greater probability of occupying a lower social position when choosing between social arrangements. This in turn could change the decision rule used by "i" to a maximin approach that focuses almost exclusively on being the worst off person in society, which would likely lead to selection of social arrangement closer to Rawls's principles in A Theory of Justice and away from a rule utilitarian conclusion.

Of course, Harsanyi would likely respond that to give the probability of being every person in society equal weight is to treat everyone impartially, and to do otherwise would contradict utilitarianism's egalitarian principles. However, this response confuses two disparate thoughts. Assigning equal moral weight to all persons in one's thinking and assigning equal probability to being each person in society when expressing preferences for a type of social arrangement under uncertainty are distinct. The former is the definitional egalitarianism advocated by utilitarianism. The latter expresses one's attitude toward gambling, and thus changing it to reflect the unequal likelihood of occupying different social positions is not a moral matter.

This paper has presented John Harsanyi's argument for rule utilitarianism and has attempted to show how the tenuousness of several central assumptions renders it an insufficient defense. Harsanyi holds that, if an individual is moral, he must be rational, and as such will regard a social system that maximizes average expected utility in society as both good and just. However, the shortcomings in his definition of a moral preference and use of his equiprobability postulate far from warrant this conclusion.

*Andrew Rennemo, MSc Philosophy and Public Policy, 2007.