Macalester Journal of Philosophy

Volume 14
Issue 1 *Spring 2005*

Article 7

5-1-2005

Utilitarianism, Game Theory and the Social Contract

Daniel Burgess

Utilitarianism, Game Theory, and the Social Contract

Daniel Burgess

I. Introduction

One of the long-standing debates in the field of ethics has to do with the ethical system of utilitarianism. Ethicists have argued for over a century about the feasibility, applicability, and the possible results of the implementation of such a system. But this wide-ranging debate over the entire system of utilitarianism often overshadows a debate which exists between utilitarians themselves. Some utilitarians feel that the best method of ethics is one which evaluates individual actions based upon their consequences. Others feel that utilitarianism should focus on finding and codifying the rules which, when universally applied, result in the greatest amount of good for the greatest number. It is this debate, the debate between act- and rule-utilitarians, which I wish to highlight and expound upon in this paper.

It is well-established by now that any attempt to universalize decisions made by purely act-utilitarian criteria may have disastrous results,[21] and it is not my purpose in this paper to rehash generally accepted arguments. However, when confronted with these critiques, many act-utilitarians rally around the fact that these overall disadvantages in utility caused by the universalization of act-utilitarian decisions are outweighed by the numerous occasions in which an act-utilitarian framework produces more desirable immediate consequences than does a rule-utilitarian

[21] Hunter, D. "Act utilitarianism and dynamic deliberation." *Erkenntnis 41* (1994): 11-12.

system.[22] Though this may be correct in certain cases, I will show that in a certain class of cases, act-utilitarian decision procedures result in an overall level of utility that is far inferior to that which would be achieved through the use of a rule-utilitarian decision procedure. Through the use of game theory and a game-theoretical model of the social contract, I will demonstrate that in the class of non-communicative games with multiple optimal equilibria, utilization of rule-utilitarian decision procedures is actually immediately superior to the use of act-utilitarian decision procedures. I will then discuss the importance of games of this type to the field of prescriptive ethics as a whole.

I. Definitions

Before I begin my analysis it is necessary to clarify a few terms. First, by *utilitarian* I mean a person who, when evaluating a choice of possible decisions or actions, believes the correct choice to be that which will result in the greatest attainment of good for the greatest number of people. In this definition I make no claims about what this good should be considered to be. In fact, for the purposes of analysis in this paper, what exactly constitutes the good is completely irrelevant. I have used the general term happiness, but replacing "happiness" with "pleasure," "satisfied preferences," or "actualized mental states" would cause no inconsistency in my argument. So long as one accepts that there is such a thing as a good which we should seek to maximize, my argument remains valid.

For act- and rule-utilitarianism, I will be using the definitions offered by Binmore. By *act-utilitarian* I

[22] Mackie, J. L. "The disutility of act-utilitarianism." *Philosophical Quarterly 23* (1973): 289-300.

mean one who "argues that each individual act should maximize the common good."[23] By *rule-utilitarian* I mean one who "argues that utilitarian principles should be applied to the rules to which we appeal when making decisions." [24] While there are interesting ethical disputes about the exact nature of the difference between rule-utilitarianism and deontological systems of ethics like that of Immanuel Kant, it is not my purpose in this paper to discuss the differences between deontology and consequentialism. I am only attempting to differentiate between act- and rule-utilitarianism.

Finally, *game theory* is defined as "the study of the ways in which *strategic interactions* among *rational players* produce *outcomes* with respect to the *preferences* (or *utilities*) of those players." [25] *Equilibrium* for a particular game is the stable combination of choices and subsequent results which follow directly from a rational evaluation of the aspects of the game based on certain underlying assumptions. The only underlying assumptions made in this paper are those constraining the decision procedures of the players of the game and the predetermined rules of the game itself. A *decision procedure* is the method by which players evaluate what constitutes a beneficial outcome of the game. The two decision procedures explored are an act-utilitarian procedure and a rule-utilitarian procedure. A *strategy* is a general theory of action stemming from the decision procedure of a player. A strategy is said to be *optimal* insofar as it produces the best possible result. With that, let the games begin!

[23] Binmore, K. *Game Theory and the Social Contract, Volume 2*. Cambridge, MA: MIT Press, 1998. p. 161.

[24] *Ibid*. p. 161.

[25] "Game theory." *The Stanford Encyclopedia of Philosophy (Online)*.

II. The Driving Game

For an example of a game which illustrates the difference between act- and rule-utilitarian decision procedures, let’s stay as simple as possible. The game described below will have but two players, each having only two possible choices. The game is far simpler than most such games in the real world, but the concepts illustrated within can be extrapolated into more complex situations quite easily. Before we begin the game, we have to make certain assumptions about its nature. These assumptions constitute the rules by which the game is played:

1) Both players must use the same decision procedure.
2) Both players are rational decision-makers.
3) Both players are able to accurately predict the outcome of a given combination of actions.
4) Both players are aware of the first three assumptions.

Tom and Jerry are our two players. Each one is driving toward the other on a road at night, and the road is just wide enough for both cars to pass one another safely. As the two cars approach one another, each driver has a choice: veer to the left or veer to the right. The range of possible outcomes for this exchange is illustrated by the decision matrix below:

		Tom	
		Left	Right
Jerry	Left	1	-5
	Right	-5	1

For the purposes of this diagram, we will measure the outcome of the event in terms of the total amount of happiness received by the two participants.[26] In this paper all outcomes will be measured in terms of total utility received by all participants of the game without reference to the individual levels of utility caused by each outcome. In this case, the pair receives 1 unit of happiness either when both participants veer to the left or when both participants veer to the right, causing the two cars to pass one another without incident. When one participant veers right as the other veers left, however, the cars crash, causing a general level of unhappiness five times more intense than the happiness experienced with an uneventful ride.[27]

III. Act-utilitarian Decision Procedures

Let us suppose that Tom is an act-utilitarian, which means that under assumption 1, so is Jerry. Tom knows that the best possible outcome will come from both drivers either choosing left or choosing right, and assumes that Jerry will come to this conclusion as well. As an act-utilitarian, what should Tom do?

The plain and simple truth is that Tom has no idea, and neither do we. This is because the game which Tom is playing has multiple equilibria. If there were only one optimal resolution to the game being played, Tom would have no trouble choosing the action

[26] Again, it is important to note that the actual type of "good" represented in this matrix is immaterial. Replace "happiness" with "satisfied preferences" and you'll get the same result.

[27] It could be argued that the decision I have chosen to analyze is not a moral one. It is not my purpose in this paper to debate what decisions do and do not constitute morality. I will simply state that it is quite easy to imagine a situation similar to the one described above that could be widely recognized as being "moral" in nature, and that I have chosen the driving game merely for its simplicity and its universality.

that would lead to that optimal resolution, and would be able to assume that Jerry, a fellow act-utilitarian, would do the same. However, in this game there are two optimal solutions, and each solution requires a different choice from Tom. The result of both players choosing right is just as good as the result of both players choosing left. If Tom chooses to go right, he faces a ½ probability that Jerry will also go right, and that a favorable outcome will result. But he also faces a ½ probability that Jerry will choose left, and that a thoroughly unfavorable outcome will come about. The same ½-½ ratio results if Tom chooses to go left. Tom has no good reason, as an act-utilitarian, to choose one over the other, and neither does Jerry. If the two players could communicate, then there would be a chance that the situation could be resolved in a mutually beneficial manner. However, this situation precludes the possibility of communication, and each participant in the game has to choose based on nothing but his rational judgment. Given that act-utilitarianism does not suggest right over left or vice versa, Tom must choose, in effect, randomly. Jerry, being a rational agent with the same decision process as Tom, makes the same choice. So the strategy which results from an act-utilitarian decision procedure is that both players choose randomly.[28] So the result for Tom is that half of the time he passes Jerry uneventfully, and half of the time the two end up exchanging insurance info on the side of the road. Clearly this is problematic. Not only is this not the best possible outcome for the game being played, but it is markedly inferior to the *actual* result of this game that we play hundreds of times every day. In this case at least, a strategy stemming from an act-

[28] Binmore, K. "Reciprocity and the social contract." *Politics, Philosophy & Economics 3* (2004): 7.

utilitarian decision procedure produces results that are far less than optimal.

IV. Rule-utilitarian Decision Procedures

Let us change things around a bit and assign a rule-utilitarian decision procedure to both players. Just as before, Tom enters the game looking for an optimal strategy given his decision procedure, and just as before, he finds one. However, the optimal strategy for the rule-utilitarian differs greatly from the optimal strategy for the act-utilitarian. Looking at the decision matrix, Tom sees that the best outcomes occur when both he and Jerry choose the same side. Included in his decision procedure is the caveat that in order for a strategy to be optimal, it must be universally applicable for this type of situation. Given this caveat, Tom reasons as follows: if I choose right, then universalizing that choice means that Jerry must choose right. And if we both choose to veer right then we miss each other and the best possible solution comes about. Tom therefore comes to the conclusion that the rule "always choose right," becomes the optimal strategy because its universalization means that Jerry must follow it as well. He is correct to reason that "always choose right" is the optimal strategy for this particular game. If this optimal strategy is followed by both players, then no accidents ever occur between the two and a far better overall outcome is achieved than that which follows from act-utilitarian reasoning.

However, there are still problems with the rule-utilitarian decision procedure that Tom utilizes. Remember that one of the restrictions to this class of games is that the two players cannot communicate. Because of this fact, Tom has to make a very suspect leap when deducing the optimal strategy for this game. He has to assume that Jerry will also come to the conclusion that “always choose right” is the rule to follow. However, given that this game has multiple equilibria, this is not always the case. A rule which may appear just as attractive to Jerry as “always choose right” is the rule “always choose left.” Jerry may reason that the universalization of this second rule would also lead to no crashes between the two players, which is the best possible outcome. What happens when the two players decide to model their strategies after different but equally optimal rules? In this case, the two players will always crash, which is an even worse outcome than that offered through an act-utilitarian analysis of the game. So in the end, unless something about the game is changed, the Nash equilibrium for the two games is exactly the same. Half of the time the two rule-utilitarian players will both choose either “always choose left” or “always choose right” as their rule, but the other half of the time they will choose opposite rules to universalize. So, just like act-utilitarians, rule-utilitarians will crash half of the time if not allowed to communicate. Something more is needed before we can call rule-utilitarianism a true optimal strategy. Tom must have a reasonable expectation that Jerry will choose the same optimal strategy before Tom’s implementation of this strategy will eliminate all crashes.

This reasonable expectation cannot be arrived at through discussion, as the game is rigged so that the two players cannot discuss what choice the other is going to make. However, it can be arrived at through other means.

V. The Social Contract

My contention is that this situation can be resolved in an optimal manner by an appeal to the social contract. The particular form of contractarianism to which I am referring is not the more traditional view ascribed to Hobbes, Locke, and Rousseau, but rather the more recent 20th-century conception espoused by John C. Harsanyi and John Rawls. Both Harsanyi and Rawls argue for an idea of the social contract as stemming from a rational agreement by the interested parties. Both Rawls and Harsanyi argued that behind a "veil of ignorance," a rational agent would agree to cede power to a government given that the methods of distribution used by the government were consistent with a rational principle. Rawls felt that this rational principle was the "maximin principle," while Harsanyi argued for a "preference utilitarianism principle."[29] Both principles have been critiqued at length in the philosophical literature, and my purpose in this paper is not to debate the merits of one principle over another. Rather, I wish to suggest that the social contract can more adequately be framed within the contexts of game theory and utilitarianism. My thinking in this area is similar to that of Ken Binmore:

> I think game theory has two major lessons for a putative science of moral behavior. The first is that a social contract can be usefully understood as the set of commonly understood conventions that allow the citizens of a society to coordinate on one of the many equilibria in their game of life. The second is that…a much wider range of behavior [is] supported as an equilibrium in repeated games than is generally thought."[30]

[29] Weirich, P. "A Game-theoretic comparison of the utilitarian and maximin rules of social choice." *Erkenntnis 28* (1988): 117-133; Harsanyi, J. C. "Rule utilitarianism and decision theory." *Erkenntnis 11* (1977): 25-53.

[30] Binmore, K. "Reciprocity and the social contract." *Politics, Philosophy & Economics 3* (2004): 5-6.

If we view the human experience as the sum total of a number of different games that we play each and every day, then the social contract functions effectively as a means of suggesting the correct outcomes of those games. When a game, such as that outlined above, has multiple equilibria, it is the purpose of the social contract to mediate between the two competing equilibria and promote one over the other. This can happen in a wide number of ways. When the game being played is of little significance in the grand scheme of things, the social contract may manifest itself in the form of social conventions or habits. However, when an improper conclusion to the game being played may jeopardize the safety or life of the participants, it becomes necessary for someone to forcibly implement the social contract. This usually comes in the form of laws issued and enforced by the government. The purpose of the government, then, is to firmly establish those rules to be followed by citizens which will result in the greatest number of fulfilled equilibria for the various *dangerous* games that we play. These rules are by no means set in stone, and may be changed if doing so would allow the participants in the game of life to fulfill more or greater equilibria.

VI. The Driving Game Revisited

Clearly, the game of driving is one in which a dangerous outcome is probable without the proper rules being followed. Given that, we can utilize this game-theoretical approach to the social contract to ensure that one of the two possible optimal equilibria in the game is universalized. In America, we universalize the rule of "always choose right." Of course, in other countries, such as Britain, the rule "always choose left," is universalized. The specific rule which is universalized is immaterial; what is important is that the participants of the game understand what rule is to be universalized and have a reasonable expectation that other players will choose to follow the same rule as they do. Assuming that the driving game mentioned above is being played in America, Tom now has a reasonable expectation that Jerry will choose to follow the state-mandated rule of "always choose right." Given that this is the rule which he can reasonably expect Jerry to universalize, Tom now chooses to universalize "always choose right" as well. The end result of the game with the addition of this contractarian assumption is that the best possible outcome will now almost always occur, and that the rule-utilitarian approach yields preferable results.

However, one could point out that the changes we have enacted in the revisited driving game are just as applicable to act-utilitarianism as to rule-utilitarianism. However, there is a fundamental problem in applying the constraints listed above to a system in which all the players are act-utilitarian. Recall that one of the advantages of a social contract listed above is that the participants of the game *have a reasonable expectation that other players will choose to follow the same rule as they do*. There is no inconsistency in applying this maxim to a game in which all participants are rule-utilitarians, because, by the very definition of rule-utilitarianism, every rule-utilitarian will follow those rules which he or she knows will produce the optimal outcome when universally applied, regardless of the circumstances of the game. An act-utilitarian has no such restriction. Perhaps the most fundamental difference between an act-utilitarian and a rule-utilitarian is that the act-utilitarian is free to break a rule which is most beneficial when universally applied, whenever the act-utilitarian deems it to be in the best interest of all people involved to break the rule.[31] Given this fundamental difference between act-utilitarianism and rule-utilitarianism, it is my contention that act-utilitarianism is incompatible with the form of the social contract outlined above because in a system in which all participants are act-utilitarian, there can be no reasonable certainty that a player will follow the rule necessary for the optimal resolution of the game, and because the players will not always choose the optimal strategy. Because no such inconsistency exists when all players utilize rule-utilitarian strategies, that system of ethics is optimal in games of this type.

It could (and has) been argued that given the simplicity of the game I have set up, there is no possible reason for an act-utilitarian to stray from conventional driving behavior because there is no reason to believe that doing so would be to anyone's benefit. To this I respond that, while this is the case in the simplest of games, it is easy to imagine a slightly more complex game for which this is not true. For example, let's put both Tom and Jerry at a bar. Both of them are convinced that they can consume vast amounts of alcohol in order to have a smashingly good time without this affecting their motor skills in any way. Needless to say, they are incorrect. A rule-utilitarian would immediately recognize that the optimal strategy in this game is to refrain from drinking. An act-utilitarian could quite easily reason that the universal rule of "don't drink and drive" is optimal when applied to everyone, but could also conclude that he is an exception due to his tolerance. Given this, in the more complex game described above, act-utilitarian Tom will

[31] See Carlson, G. "Plans, expectations, and act-utilitarian distrust." *Philosophical Studies 36* (1979): 295-300; Freedman, B. "A meta-ethics for professional morality." *Ethics 89* (1978): 1-19; Lyons, D. *Forms and limits of utilitarianism.* Oxford, UK: Clarendon Press, 1965.

not always follow the optimal strategy, and can have no certainty that act-utilitarian Jerry will follow the optimal strategy either. In this case, then, the social contract as defined above and a prescriptive act-utilitarian system of ethics are incompatible.

VII. Objections

The most obvious objection to the situation present above deals with the restrictions placed upon the actors. Specifically, the driving game as was framed above greatly confined the players' knowledge and opportunities for cooperation by prohibiting them from communicating. In real life, one might argue, we are not prevented from communicating from those with whom we play games. We can talk to them and discuss which set of choices might result in the best overall outcome. Therefore, while the driving example above might be illustrative of a given case, this case is rare and contrived enough that the arguments applicable in this situation are not applicable to the entirety of the range of choices available to members of a society.

I am willing to concede that the above situation is a particular case, but it belongs to a class of cases which I feel have a great deal of importance in the field of prescriptive ethics. Because of the vastness of our cities, the increasing influence of mass media, and the growing world population, it is becoming more and more difficult every day to have meaningful interactions with those with whom we do not already have a prior relationship. The important point here is not that the above class of games relates only to those cases in which it is *impossible* to communicate with the others playing, but also to those cases in which it is *infeasible* to do so. Every day we are confronted by a multitude of situations in which we could take the time to communicate with other actors in order to mutually consent to the best possible combination of actions, but we choose not to because of constraints upon our time and resources. The game of driving is but one of the situations in which an optimal outcome could be arrived at through cooperative deliberation; however, in this game and in many others we defer to rules to save time, because no one wants to spend their driving time shouting out directions to every other person on the road. Life is filled with situations of this type, and rule-utilitarianism, when conjoined with a game-theoretic model of the social contract, can ensure that the optimal equilibria to the many impersonal games that we play everyday will occur.

VIII. Conclusion

My goal in this paper was to offer a selective refutation of the argument that application of an act-utilitarian decision procedure to an individual situation will necessarily result in an optimal result. It is my belief that a more wide-sweeping refutation of this argument is possible given unlimited space in which to present ideas; however, given the restricted length of this paper, I chose to focus upon a single class of games to which this argument cannot apply. Through an analysis of the specific class of non-communicative, multiple-equilibria games, it is apparent that oftentimes an act-utilitarian decision procedure is markedly inferior to a rule-utilitarian decision procedure in terms of the strategies suggested by each approach. However, in order for rule-utilitarianism to truly prescribe an optimal strategy for the player of the game, a neo-contractarian position such as the one offered by Binmore is necessary to ensure the preferability of one of a number of equally optimal equilibria. Applied to the "game of life," it is my belief that further implementation of this approach to prescriptive ethics may shed light upon and serve to improve those rules which best promote the happiness of society as a whole.