Multiple Moralities:

A Game Theoretic Examination of Indirect Utilitarianism

Paul Studtmann

Shyam Gouri-Suresh

**Abstract**

In this paper, we provide a game theoretic examination of indirect utilitarianism by comparing the expected payoffs of attempts to apply a deontological principle and a utilitarian principle within the context of the Prisoner’s Dilemma (PD). Although many of the best-known utilitarians and consequentialists have accepted some indirect form of their respective views, the results in this paper suggest that they have been overly quick to dismiss altogether the benefits of directly enacting utilitarian principles. We show that for infallible moral agents, what we call ‘non-autonomous agents’, direct utilitarianism dominates indirect utilitarianism via deontology in terms of achieving the maximized utilitarian outcome, but only in underlying games where the maximized utilitarian outcome involves unequal payoffs. In other situations, indirect utilitarianism implemented through Kantian deontology either ties or dominates direct utilitarianism in terms of achieving the maximized utilitarian outcome. We also examine the two different moralities on the assumption that fallibility, which is a form of autonomy, is an aspect of moral agency by introducing *Endogenized Morality Models (EMM’s)*. We believe that just as indirect utilitarians worry about the cost of applying moral principles, so too they should worry about the fact that humans have both pro-social and materialistically selfish motivations and hence are fallible moral agents. We show that there are conditions under which fallible autonomous utilitarians achieve higher expected material and psychic payoffs than fallible autonomous deontologists and conditions under which they do not.

According to indirect utilitarians, agents do not maximize happiness by computing the consequences of an action and trying to act accordingly. Hence, agents ought to adopt a decision procedure that does not involve trying to apply the utilitarian principle. Judging by the number of well-known indirect utilitarians and consequentialists the arguments for indirect utilitarianism have been very successful. The list includes J.S. Mill, John Austin, Henry Sidgwick, J.J.C. Smart, R.M. Hare, David Brink, G.E. Moore, Peter Railton, Derek Parfit, and Philip Pettit. (Wiland 2007) Despite this success, however, the arguments have largely been informal. This is not surprising. To compare formally the effect of applying a principle of utility and some other principle, for instance a deontological principle, one would need a unified mathematical way of characterizing agents who can apply moral principles in various contexts. And such a unified mathematical theory is difficult to find.

In the standard game theoretic approach, moral preferences are represented by exogenous parameters that modify an agent’s utility function. Several such parameters have been introduced. Although successful in many respects, the parameter approach makes a unified comparison of different moralities difficult, since it would require a unified treatment of the exogenous parameters. In this paper, we present an alternative approach to modelling moral preferences, one that is related to the parameterized approach but that does not ultimately require exogenous parameters and so allows for a unified game-theoretic treatment of moral agency. The approach that we adopt occurs in Studtmann and Gouri-Suresh (2020) and Gouri-Suresh and Studtmann (2022). The approach involves introducing a type of model that we call an *Endogenized Morality Model* (EMM). As we show, EMM’s allow for a unified game-theoretic comparison of the expected payoffs that result from agents applying a utilitarian principle and agents applying a deontological principle.

EMM’s are constructed in two steps. First, a morality is represented directly within game theory by way of a transformation of an original game. The transformation can be given in two ways. The most intuitive way is to require that the transformation reflects outcomes that result when the world is the way it ought to be according to a principle that defines the moral theory. We call the transformation of PD according to a deontological principle *Deontology*. In *Deontology*, an agent receives what she would have received were everyone to act as she does. In this way, she receives those payoffs that would occur were everyone acting according to the same principle. *Deontology* thus captures the underlying Kantian intuition that the moral law must be universalizable. We call the transformation of PD according to a utilitarian principle *Utilitarianism*. In *Utilitarianism*, an agent receives the average of the total payoffs of an outcome. The outcomes thus reflect the way that utilitarians think payoffs ought to be distributed. Finally, we assume that the identity transformation yields a morality of materialistic selfishness.

In addition to the intuitive derivation of *Deontology* and *Utilitarianism*, it is possible to derive them from well-known parameterizations of the different preferences associated with them. Alger and Weibull (2014, 2015, 2018) introduce a morality parameter, k, that represents what they call an agent’s *Kantian preferences*, which are an agent’s preference for outcomes in which everyone acts as she does. When , PD with the Alger-Weibull utility function reduces to *Deontology*. And Marc Levine (1998) introduces an altruism parameter, , that represents an agent’s altruistic preferences for her counterparty’s outcome. When , PD with Levine’s utility function reduces to *Utilitarianism*. Hence, each morality represents an agent that has the most extreme possible preference for the type of outcome associated with the respective moralities.

In the second step to constructing an EMM, a morality space is defined for each moral agent which describes the set of alternative moral principles that the moral agent can choose amongst. Let denote the set of alternative moral principles available to each moral agent. In an EMM, agents choose not just their action strategy of cooperate or defect from the action strategy space, , but also their moral principle from the morality space . The transformed strategy space, , available to each moral agent is the Cartesian product of the morality space and the action strategy space, .

Although the approach is general and can be applied to any standard normal form non-cooperative game, in this paper we restrict our attention to transformations of PD, and we provide a comparison of the outcomes of two types of agents who apply two different moral principles to PD. We use the phrase ‘non-autonomous agent’ to describe an agent who plays according to a transformed matrix. We use the phrase ‘autonomous agent’ to describe an agent who has a choice between playing according to more than one morality. There is no constraint on the number or types of morality an autonomous agent can choose from. An EMM can be constructed, for instance, that requires an agent to choose between acting as a utilitarian and acting as a deontologist. In this paper, however, we restrict our attention to EMM’s in which the choice is between acting according to a morality of materialistic selfishness and acting according to some pro-social morality. Just as real-world agents must decide between acting in a materialistically selfish way and acting according to a moral principle, autonomous agents as we model them in this paper must choose not only whether to defect or cooperate in a PD but also whether to do so while acting according to materialistically selfish motivations or while acting in accordance with a moral principle. As we show, the struggle between selfishness and morality leads to fallible moral agents who are less cooperative than infallible moral agents but more cooperative than infallible selfish agents typically studied in game theory.

The EMM’s we discuss in this paper do not include the cost involved when an agent tries to apply a moral principle. We thus sidestep one argument that has been raised against direct Utilitarianism. According to the objection, the cost of applying a utilitarian principle offsets whatever benefit might result from applying it. Agents thus produce more happiness when they don’t try to apply a utilitarian principle. (Wiland, 2007) Although we do not address this issue in this paper, it is possible to augment the moralities we examine with a variable that represents the cost of applying the moral principle associated with it. The resulting models would allow for a detailed comparison of the expected outcomes of applying different moral principles on the assumption that the costs of applying the different principles differ.

We focus on utilitarianism and deontology in this paper because of the historical and theoretical significance of the two moral philosophies. We want to stress, however, that the approach taken here can be extended to other moral philosophies, for instance a moral philosophy based on a principle of empathy, or one based on Rawlsian preferences. (See Gouri-Suresh and Studtmann (2022) for a discussion.) We focus on PD because of its importance in modeling social dilemmas and the relevance of social dilemmas for discussions about morality. (See Kuhn 2004 for an overview and discussion.) Of course, PD isn’t the only game that is important to morality. Indeed, the difference between utilitarianism and deontology as applied to coordination games, anti-coordination games, and bargaining games, among others, are all important and fascinating. As will become apparent, however, the details of the comparison of the two moralities as applied to PD are complex enough that it is best to reserve for another occasion a broader examination.

Modelling morality and moral autonomy without parameters allows for a comparison of the expected payoffs for different moralities in both their autonomous and non-autonomous forms. It turns out that the results are mixed. We show that in the case of PD there are conditions under which acting like a utilitarian, whether autonomously or non-autonomously, leads to worse outcomes than acting according to a deontological principle and conditions under which it leads to better outcomes. The mathematical results thus support a limited rather than a global form of indirect utilitarianism.

The rest of the paper is structured as follows.

In section I, we discuss our game theoretic representation of *Utilitarianism* and *Deontology*. We show that each can be derived either from a transformation of PD that reflects a moral principle or by extremizing a parameter that represents deontological or utilitarian moral preferences.

In section II, we compare the expected payoffs for players playing the moralities we define in section I. We limit our examination of the moralities to interactions in which both agents are utilitarians or both are deontologists. We do so partly for the sake of simplicity and partly to impose a kind of deontological constraint on the examination: By requiring both agents to play by the same morality, we examine the extent to which the morality is effective in providing a single moral principle that everyone obeys.

In section III, we introduce the models of autonomous moral agents. Unlike non-autonomous agents, most humans struggle between acting in accordance with a moral principle and acting selfishly. We model this struggle directly by modeling agents in a PD as having two choices: whether to cooperate or not, and whether to do so while playing a morality or while playing the original PD. We discuss the two matrices that are the combination of PD and *Deontology* and PD and *Utilitarianism*.

In section IV, we compare the expected payoffs of autonomous utilitarians and deontologists.

Before beginning, we note that the mathematical details, especially in sections II and IV, are rather involved. In order to assist the reader, we have summarized the results in tables at the end of each of those sections. A reader who wants to follow the argument without wading into the mathematical details should be able to do so by reading the non-mathematical commentary and inspecting the tables. For the sake of presentation, we have not provided any derivations of the results in the body of the paper. But we have included an appendix that contains Mathematica notebooks that demonstrate those results that are central to the arguments being made and that are not easily derivable from an inspection of the games.

Section I

*Deontology* and *Utilitarianism*

We begin our discussion of moralities with the following standard representation of PD with two players, each of whom can choose either to cooperate or to defect . The payoffs associated with this standard PD are given in Table 1. In this game, if two cooperators interact, each gets the payoff , the “reward for mutual cooperation.” If a cooperator meets a defector the cooperator gets , the “sucker’s payoff,” while the defector gets , the “temptation of defection.” If two defectors interact, each obtains the payoff , the “punishment” of mutual defection. The game is a prisoner’s dilemma if .

Table 1: Prisoner's Dilemma (PD)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Player 2 | |
|  |  |  |  |
| Player 1 |  |  |  |
|  |  |  |

In order to capture the content of various moralities, one can either transform PD by a rule that reflects the principle of the morality in question or extremize a parameter that represents the moral preferences associated with the morality.

Kantian deontology urges agents to consider whether their actions are universalizable. Although Kant’s formulation of his deontology is by way of his categorical imperative, a common-sense rationale for his view can be given by appeal to the simple question that is used to motivate moral behavior: ‘what if everyone were to do that?’. The morality that we call *Deontology* results from transforming PD by a rule that in effect forces agents to live with the results of everyone acting like she does. According to the transformed game, an agent receives as a payoff what she would have received in the original PD had both players acted in the way that she is acting. If a player cooperates, she receives what she would have received in PD had both players cooperated, and if she defects, she receives what she would have received in the original PD had both players defected.

If both agents universalize their actions in this way, they play the following game.

Table 2: Deontology

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Player 2 | |
|  |  |  |  |
| Player 1 |  |  |  |
|  |  |  |

This intuitive route to deontology can be augmented by a route that goes through a parameterization. Alger and Weibull (2014) introduce what they call the *morality parameter*, k, which represents the extent of an agent’s preference for outcomes in which both players act the same and hence conform to Kant’s categorical imperative. Here is their definition of what they call a *homo moralis*, i.e., an agent with some possibly zero amount of preference for such outcomes.

An individual is a homo moralis if her utility function is of the form uκ(x,y) = (1 − κ) · π(x, y) + κ · π(x, x), for some κ ∈ [0 1], her degree of morality.

The following is PD with the Alger-Weibull utility function in the context of pure strategies.

Table 3: Parameterized Deontology

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Player 2 | |
|  |  |  |  |
| Player 1 |  |  |  |
|  |  |  |

When , *Parameterized Deontology* reduces to PD, and when , it reduces to *Deontology*. Hence, two agents playing *Deontology* have the maximum possible preference for outcomes in which each agent plays the same action. They are Kantian saints.

Utilitarianism is the view that one ought to maximize happiness impartially considered. A simple rule for distributing payoffs that reflects the utilitarian principle is to divide the total payoffs by the number of players and give each player the same amount. With a two-player PD, this results in the following game.

Table 4: Utilitarianism

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Player 2 | |
|  |  |  |  |
| Player 1 |  |  |  |
|  |  |  |

As with *Deontology*, there is a route to *Utilitarianism* that goes through a parameterization. Marc Levine introduces a parameter, , that represents the level of an agent’s altruism. In the n-person game, with players , player receives a direct utility of . Player has a coefficient of altruism . And player i’s altruistic preferences are dependent on his attitude about the other player’s altruism, which is represented by . We present here Levine’s Utility function with a rescaling factor.

Gouri-Suresh and Studtmann (2022) have argued that moral principles should be counterparty independent, by which they mean that a moral principle ought to be applied independent of the other person’s morality. With such an assumption, . Hence, in a 2-player PD, Levine’s utility function yields the following payoff matrix.

Table 5: Parameterized Utilitarianism

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Player 2 | |
|  |  |  |  |
| Player 1 |  |  |  |
|  |  |  |

When , this matrix reduces to PD. And when , this matrix reduces to *Utilitarianism*. Hence, two agents playing *Utilitarianism* have the maximum possible altruistic preferences. They are utilitarian saints.

*Deontology* and *Utilitarianism*, then, can be considered in two different lights. First, they result from transformations of PD which reflect an underlying moral principle. Second, they result from parameterizations of the moral preferences associated with different moralities when the parameter in question is extremized. Whichever way one views them, they represent what game theorists typically call *psychic* as opposed to *material* payoffs. An agent who plays *Deontology*, for example, has the most extreme possible deontological psychic payoffs. Such an agent, therefore, perfectly exemplifies a moral principle based on deontology. Although the distinction between psychic and material payoffs is standard, it is important to stress that in the models in this paper, the distinction is really between moral psychic payoffs and non-moral psychic payoffs. Because the payoffs in the original PD are utilities, they are already in some sense psychic as opposed to material payoffs. The payoffs in *Utilitarianism* and *Deontology* are transformations of the original payoffs by a moral principle and so represent distinctively moral psychic payoffs.[[1]](#footnote-1)

With *Deontology* and *Utilitarianism*, we have models of agents who perfectly exemplify the motivations of their respective moral philosophies. Of course, this kind of agent is rare if not entirely non-existent. Most moral agents struggle with the choice as to whether to be selfish or moral. In the models we discuss in section III, agents choose between material selfishness, which is represented by the original PD, and one of the two moralities. Before addressing autonomous agents, however, we first compare the two different moralities as exemplified by agents who have no choice but to play according to their rules.

Section II

*A Comparison of Non-Autonomous Agents*

The behavior of non-autonomous deontological agents in PD is easy to describe. In *Deontology*, when , the pure strategy of cooperation is the only Nash equilibrium. Hence, the expected psychic and material payoffs for deontological non-autonomous agents equal . It is a standard claim that deontology is committed to a form of absolutism. If one restricts oneself to PD, then deontology is indeed absolutist. A non-autonomous deontological agent always cooperates in a PD. Although our focus here is PD, it is worth mentioning that under the orderings that generate the Stag-Hunt and Hawk-Dove *Deontology* has the pure strategy of Stag and Dove, respectively, as the only Nash equilibria. Hence, *Deontology* yields absolute imperatives for cooperation, payoff-dominant coordination, and peaceful co-existence, all of which are very reasonable imperatives for well-functioning human groups.

*Utilitarianism*, on the other hand, does not entail absolute imperatives. The solution space for *Utilitarianism* divides into three regions:

1. ;
2. ;

.

When holds, *Utilitarianism* becomes an anti-coordination game. Hence, there are two asymmetric pure strategy equilibria as well as a symmetric mixed strategy equilibrium. In the asymmetric equilibria, the total expected material and psychic payoff equals . Because , (. Hence, the total expected psychic payoffs and the total expected material payoffs in the asymmetric equilibrium in *Utilitarianism* under condition (1) are greater than or equal to the total expected psychic and material payoffs in *Deontology* under the same condition.

In the symmetric mixed strategy for *Utilitarianism*, each agent cooperates with the following probability:

The expected payoff, EP, for each agent is given by the following equation:

It can be shown that in any symmetric equilibrium for *Utilitarianism* the expected psychic and material payoffs are the same. Hence, this equation expresses both the expected material and psychic payoffs for the equilibrium. It can be shown that when holds, . Hence, when holds, non-autonomous agents in the symmetric mixed-strategy equilibrium achieve higher expected material and psychic payoffs than agents playing *Deontology*. Moreover, unlike in the asymmetric equilibria, utilitarianagents in the symmetric equilibrium achieve the same expected payoffs. Hence, when condition (1) holds, both the asymmetric and symmetric equilibria have higher expected payoffs than the equilibrium in *Deontology*.

When holds, the only strategy in equilibrium in *Utilitarianism* is the pure strategy of cooperate. Hence, when holds, utilitarian agents achieve the same outcomes as non-autonomous deontologists.

When holds, *Utilitarianism* is a coordination game. Hence, there are two symmetric pure-strategy equilibria and one symmetric mixed strategy equilibrium. The pure strategy equilibrium of cooperation in *Utilitarianism* has expected payoff of R. Hence, it yields the same material and psychic payoffs as *Deontology*. The symmetric pure strategy equilibrium of defect has expected psychic and material payoffs equal to . Because , in such an equilibrium *Utilitarianism* has lower expected material and psychic payoffs than *Deontology*. The expected material and psychic payoffs for the symmetric equilibrium are given by the equation for EP above. It can be shown that when holds, . Hence, when (3) holds, in two of the three Nash equilibria, *Utilitarianism* has (much) lower expected material and psychic payoffs than *Deontology*.

It may help to summarize the results so far in the language of winners and losers. If we restrict ourselves to non-autonomous agents, and if we use as our metric expected payoffs of two agents playing the same morality, then under *Utilitarianism* wins; under *Utilitarianism* and *Deontology* tie; and under *Utilitarianism* sometimes (in one Nash equilibrium) ties *Deontology* and sometimes (in two Nash equilibria) loses to it badly. It is worth noting that these results provide a general rule for deciding when acting as a non-autonomous utilitarian is a better bet than acting as a non-autonomous deontologist. The decision depends on the total value of the asymmetric outcomes relative to the total values of the symmetric outcomes. When the total value of the asymmetric outcomes is low, more specifically when , then deontology is a better bet than utilitarianism. But as the total expected value of the asymmetric outcome increases, utilitarianism at first becomes no riskier than deontology and then, when , becomes a better bet than deontology.

It is worth noting that the condition under which *Utilitarianism* does better than *Deontology* is precisely the one that utilitarians are (in)famous for endorsing, namely the condition in which unequal outcomes in the underlying game yield the greatest overall happiness. It is thus, perhaps predictably, in precisely that condition that generates the theoretical suspicion that utilitarianism is not the correct moral theory that utilitarianism is best suited for achieving its own ends. But perhaps less predictably, *Utilitarianism* has a way of achieving some of the benefits that such a condition makes possible by way of a symmetric mixed strategy equilibrium in which both agents achieve the same expected payoffs. When , the symmetric mixed strategy equilibrium has expected payoffs greater than R and so greater than what deontology could deliver. Although greater than , however, the total expected payoffs for the strategy are less than and hence less than the total payoffs in the asymmetric equilibria. Utilitarians in a symmetric equilibrium can achieve higher outcomes than deontology; but to maximize total expected payoffs they must occupy an asymmetric equilibrium.

It is useful to consider these results with Rawls’s veil of ignorance in mind. Suppose someone had to choose whether to play *Deontology* or *Utilitarianism* but did not know what the values of T, R, P, and S would be, and she also did not know whether she would cooperate or defect if she ended up in an asymmetric equilibrium. Her choice would depend in part on whether she employed maxi-min reasoning. If she did, she would choose *Deontology*. In so choosing*,* she would choose a morality with an absolute imperative to cooperate (as well as absolute imperatives to play dove and stag), and she would be certain that her expected outcome would equal R. In choosing utilitarianism, on the other hand, she would choose a moral philosophy with no absolute imperatives and no guarantee of a good outcome. But she would also be choosing the possibility of getting lucky. If it were the case that her own payoffs would be better than they would be in *Deontology* if either the equilibrium was symmetric, or she were the defector in the asymmetric equilibrium.

The mathematical facts discussed so far are summarized in the following two tables. The first gives the total expected material and psychic payoffs, and , for the two types of agents under three different conditions. The second gives the Nash equilibria of the same two types of agents under the same conditions: is the symmetric pure strategy of cooperation, is the symmetric pure strategy of defection, and are the asymmetric pure strategies of cooperate defect and defect cooperate, and is the symmetric mixed strategy that is a mixture of cooperate and defect.

*Table 6(b): Nash Equilibria for Non-Autonomous Agents*

|  |  |  |
| --- | --- | --- |
|  | *Non-Autonomous*  *Utilitarian* | *Non-Autonomous*  *Deontologist* |
|  |  |  |
|  |  |  |
|  |  |  |

*Table 6(a): Total Expected Payoffs for Non-Autonomous Agents*

|  |  |  |
| --- | --- | --- |
|  | *Non-Autonomous*  *Utilitarian* | *Non-Autonomous Deontologist* |
|  |  |  |
|  |  |  |
|  |  |  |

Section III

*Autonomous Agents*

Central to Kant’s understanding of autonomy is the capacity to obey a law that reason has given to itself. (Kant, 2019) An autonomous will can overcome the influence of the passions, society’s expectations, or any other factor that makes a will subject to a force outside itself and can choose freely to obey the Categorical Imperative. Kant obviously did not develop his theory of autonomy with game theory in mind. But the kind of capacity he envisions can be understood game theoretically by allowing agents a choice to act according to different moralities. To capture the struggle involved in human actions, it is useful to require agents to choose between a morality of materialistic selfishness and some other morality. The resulting models describe fallible moral agents whose fallibility lies in their having to choose not just how to act but also whether they will be motivated by their moral sense or their selfish interests, whether they will choose to listen to the better angels of their natures or the worse.

In an EMM, moral autonomy of this sort is modeled directly without recourse to parameters. Although Kant thought that the reason of an autonomous agent would be directed at his own version of the moral law, EMM’s expand the possible autonomous choices an agent can make to include moralities other than deontology. The following matrices are the models for, *Autonomous Deontology PD (ADPD)*, and *Autonomous Utilitarianism PD (AUPD)*, respectively. In each model, agents must make two choices, whether to cooperate or defect and whether to play according to the morality in question or to play according to the morality of materialistic selfishness.

*Table 7: Autonomous Deontology PD (ADPD)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

*Table 8: Autonomous Utilitarianism PD (AUPD)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Section IV

*A Comparison of Autonomous Utilitarianism and Deontology*

Section IV (a) – Autonomous Deontology

We begin the comparison of autonomous utilitarianism and deontology with *ADPD.* It is not hard to see that if , strictly dominates and strictly dominates . So, APDP reduces to the following matrix:

*Table 9: ADPD Reduced*

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

This game is an anti-coordination game. So, unlike in *Deontology*, the pure strategy of (moral) cooperation is not in equilibrium. Rather, there are two asymmetric pure strategy equilibria – and – and one symmetrical mixed strategy equilibrium, which is a mixture of and . Unlike their non-autonomous counterparts, autonomous deontologists are pulled in both a pro-social and a selfish direction: sometimes they cooperate and sometimes they defect. This is intuitive. When an agent has no choice but to act as a deontologist, she can’t help but cooperate. The moral law determines entirely her will. But when given a choice between acting morally and acting selfishly, sometimes she chooses to be moral and sometimes she chooses to be selfish. In this way, autonomous deontological agents of the sort we are considering are morally fallible.

In the asymmetric pure strategy equilibria, the psychic payoffs equal either R or T and the material payoffs equal either S or T. In the mixed strategy equilibrium for *ADPD*, the probability of cooperation is given by the following equation.

This probability function has intuitive consequences. As the temptation, T, to defect increases, so too does the probability of defection. And as the stakes of the dilemma, which are measured by the difference between R and P, increase, so too does the probability of cooperation. The expected psychic and material payoffs for the equilibrium, EPP and EMP, are given by the following equations.

Section IV (b) --Autonomous Utilitarianism

The behavior of autonomous utilitarians is more complex than the behavior of autonomous deontologists. Unlike ADPD, AUPD does not reduce to a 2 x 2 matrix but rather to the following 3 x 3 matrix.

*Table 10: AUPD Reduced*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

The solution space to this game divides into two regions:

1. ; and

When (1) holds, the mixed strategy and is in equilibrium as are all the pure and mixed strategy equilibria that can be constructed from and . If we distinguish between asymmetric equilibria that differ only by label, there are six possible equilibria.

In this region of the solution space, five equilibria – (ii), (iii), (iv), (v), and (vi) – consist of players defecting with probability equal to 1. In these equilibria, the expected material and psychic payoffs equal . In the mixed strategy equilibrium, (i), the probability of playing is given by the following equation.

The expected psychic and material payoffs of playing the mixed strategy are given by the following two equations.

When (2) holds, the following strategies are in equilibrium.

In the two pure strategy equilibria, (ii) and (iii), the expected psychic payoff equals either or , and the expected material payoff equals either or . The expected material and psychic payoffs for the mixed strategy equilibrium, (i), are the same as in the mixed strategy in the previous condition and so are given by the above two equations.

Section IV (c) – A Comparison

Because of the variability in the material and psychic payoffs for both utilitarian and deontological autonomous agents, a comparison of the two involves somewhat complex conditions. One part of one condition is complex enough that we have not given it in the body of the paper but instead display it in the appendix. The general take away is that with respect to both psychic and material payoffs there are conditions under which an agent does better as an autonomous utilitarian and conditions under which she does not. So, there is no *a priori* way of saying which strategy should be used. Moreover, unlike in the case of autonomous agents, maxim-min reasoning behind a veil would not lead an agent to prefer playing the deontological version of the game.

One can begin the comparison with the psychic payoffs for the asymmetric pure strategy equilibria in both ADPD and AUPD. In ADPD, the total psychic payoffs in the asymmetric equilibria equal . In AUPD, the total psychic payoffs equal . Hence, if , the total psychic payoffs in the asymmetric pure strategy equilibria for AUPD are greater than the total psychic payoffs for ADPD. If , then the total psychic payoffs in the asymmetric pure strategy equilibria are less than the total psychic payoffs in ADPD. In addition to these asymmetric pure strategy equilibria, AUPD has three equilibria when in which the strategies all involve defection. In these equilibria, the total expected psychic payoff equals . Because , in these equilibria the total psychic payoffs for AUPD are less than the total psychic payoffs for the asymmetric equilibria in ADPD.

The total material payoffs in the asymmetric equilibria for both AUPD and ADPD equal and so are equal.

The comparison of the mixed strategies is more complex. Under the following conditions, the total expected psychic payoff for the mixed strategy equilibrium in AUPD is greater than the total expected psychic payoff for the mixed strategy equilibrium in ADPD.

1. , or
2. and

Under the following conditions, the total expected material payoff for the symmetric mixed strategy equilibrium in AUPD is greater than the total expected material payoff for the symmetric mixed strategy equilibrium in ADPD (X is given in the appendix).

1. and
   1. , or
   2. and , or
2. , and
   1. , or
   2. ,

These results allow a useful contrast between the autonomous and non-autonomous versions of utilitarianism and deontology. In the non-autonomous version of *Utilitarianism*, when , it is possible for one player in an asymmetric equilibrium to have a material payoff equal to S and one player to have a payoff equal to T. Because , one player does better in that equilibrium than she would in *Deontology* and one player does worse. As a result, if someone did not know whether would be greater than , and if she also did not know whether she would play cooperate or defect in an asymmetric equilibrium, then if she were committed to maxi-min reasoning she would choose *Deontology*. Someone who is not committed to maxi-min reasoning but instead employed a standard utilitarian calculus, on the other hand, might choose *Utilitarianism*. In the autonomous versions of the moralities, such a difference no longer exists: In both AUPD and ADPD there is an asymmetric pure-strategy equilibrium in which one agent receives material payoffs equal to and one agent receives a material payoff equal to . Hence, a deontological autonomous agent may end up defecting in an asymmetric equilibrium and receive as a material payoff. Playing *Autonomous Deontology* does not guarantee a material payoff equal to R. Moreover, the psychic payoffs in the asymmetric equilibria for *Utilitarianism* can be both greater or less than the psychic payoffs in the asymmetric equilibria for *Deontology*. Hence, maxi-min reasoning behind a veil does not necessarily lead to a choice of the deontological version of the autonomous game. Moral fallibility alters the reasoning behind a Rawlsian veil.

The following table gives the strategies in equilibrium in AUPD and ADPD, where is the strategy of moral cooperation, is the strategy of moral defection, and is the strategy of non-moral defection.

*Table 11(a): Nash Equilibria for Autonomous Agents*

|  |  |  |
| --- | --- | --- |
|  | *Autonomous Utilitarian* | *Autonomous Deontologist* |
|  |  |  |
|  |  |  |

Table 11(b) – Total Expected Payoffs for Autonomous Agents

|  |  |  |
| --- | --- | --- |
|  | *Autonomous Utilitarian* | *Autonomous Deontologist* |
| *Asymmetric Pure Strategy Equilibria* | , or | , or |
| *Symmetric Mixed Strategy Equilibria* |  |  |
| *Asymmetric Pure Strategy Equilibria* |  |  |
| *Symmetric Pure Strategy Equilibria* |  | X |
| *Symmetric Mixed Strategy Equilibria* |  |  |

*Conclusion*

In this paper, we have provided a game theoretic analysis of both different moralities and different types of autonomous agent. Such an examination allows for a mathematical evaluation of claims that are central to the debate concerning indirect utilitarianism. A comparison of the expected outcomes that result from applying utilitarian or deontological principles has considerable nuance. For agents with infallible morality, i.e., non-autonomous agents, in a PD direct utilitarianism unambiguously dominates indirect utilitarianism via deontology in terms of achieving the maximized utilitarian outcome, but only in underlying games where the maximized utilitarian outcome involves unequal payoffs. In other situations, indirect utilitarianism implemented through Kantian deontology either ties or dominates direct utilitarianism in terms of achieving the maximized utilitarian outcome. We have also examined the two different moralities on the assumption that fallibility, which is a form of autonomy, is an aspect of moral agency by introducing *Endogenized Morality Models (EMM’s)*. We believe that just as indirect utilitarians worry about the cost of applying moral principles, so too they should worry about the fact that humans have both pro-social and materialistically selfish motivations and hence are fallible moral agents. We have shown that there are conditions under which fallible autonomous utilitarians achieve higher expected material and psychic payoffs than fallible autonomous deontologists and conditions under which they do not. Although many of the best-known utilitarians and consequentialists have accepted some indirect form of their respective views, the results in this paper suggest that they have been overly quick to dismiss altogether the benefits of directly enacting utilitarian principles.

We conclude by reiterating the limited nature of the results presented here. We have examined utilitarianism and deontology only as applied to PD. There are of course many other games and other moral principles. A thorough investigation would require applying all the various intuitively plausible moral principles in both their non-autonomous and autonomous forms to a wide range of games. We hope to have demonstrated in this paper only that such an examination is not just possible but that it promises to deepen our understanding of moral philosophy.

**Works Cited**

Alger, Ingela, and Jörgen W. Weibull. “Evolution and Kantian Morality.” *Games and Economic Behavior*, vol. 98, 2016, pp. 56–67., https://doi.org/10.1016/j.geb.2016.05.006.

Alger, Ingela, and Jörgen W. Weibull. “Evolutionary Models of Preference Formation.” *Annual Review of Economics*, vol. 11, no. 1, 2019, pp. 329–354., https://doi.org/10.1146/annurev-economics-080218-030255.

Gouri Suresh, Shyam, and Paul Studtmann. “Angels and Devils on Our Shoulders: Modeling Moral Agency.” *SSRN Electronic Journal*, 2022, https://doi.org/10.2139/ssrn.4028754.

“Homo Moralis--Preference Evolution under Incomplete Information and Assortative Matching.” *Econometrica*, vol. 81, no. 6, 2013, pp. 2269–2302., https://doi.org/10.3982/ecta10637.

Kant, Immanuel, and Thomas Kingsmill Abbott. *Groundwork of the Metaphysics of Morals*. Compass Circle, 2019.

Kuhn, Steven T. “Reflections on Ethics and Game Theory.” *Synthese*, vol. 141, no. 1, 2004, pp. 1–44., https://doi.org/10.1023/b:synt.0000035846.91195.cb.

Levine, David K. “Modeling Altruism and Spitefulness in Experiments.” *Review of Economic Dynamics*, vol. 1, no. 3, 1998, pp. 593–622., https://doi.org/10.1006/redy.1998.0023.

Studtmann, Paul, and Shyam Gouri Suresh. “Universalizing and the We: Endogenous Game Theoretic Deontology.” *Economics and Philosophy*, vol. 37, no. 2, 2020, pp. 244–259., https://doi.org/10.1017/s0266267120000279.

Wiland, Eric. “How Indirect Can Indirect Utilitarianism Be?” *Philosophy and Phenomenological Research*, vol. 74, no. 2, 2007, pp. 275–301., https://doi.org/10.1111/j.1933-1592.2007.00018.x.

1. Understanding the distinction in this way is important. According to the standard articulation of utilitarianism it is happiness that should be maximized. Because happiness is associated with psychic payoffs, one ought to maximize psychic payoffs. But which ones? The moral or the non-moral? As we have defined Utilitarianism, it is the non-moral psychic payoffs that ought to be distributed according to the utilitarian principle. One might object that according to Utilitarianism the moral psychic payoffs ought to be distributed according to the utilitarian principle. Such an objection, however, would require utilitarianism to be circular. The circularity can be seen both non-mathematically and mathematically. If utilitarianism is an attempt to define how one morally ought to behave, then defining it in terms of maximizing moral psychic payoffs includes in the definition the very concept that is being defined. Mathematically, the circularity leads to undefined payoffs. To see the problem, suppose player 1 is a utilitarian with a moral psychic payoff equal to . And suppose that player 2 is a utilitarian with a moral psychic payoff equal to Y. Because player 2 is a utilitarian who maximizes moral psychic payoffs, her moral psychic payoff must equal , which implies that . So, player 2’s moral psychic payoff depends on player 1’s moral psychic payoff. But, because player 1 is a utilitarian, player 1’s moral psychic payoff also similarly depends on player 2’s moral psychic payoff. And so, an attempt to define the moral psychic payoffs in terms of moral psychic payoffs leads to circularity. [↑](#footnote-ref-1)