Heteronomy v. Autonomy

Shyam Gouri Suresh* Paul Studtmann†

July 16, 2024

Preliminary and Incomplete.
Please do not cite or circulate.

Abstract

Kant distinguishes between autonomous and heteronomous agents. Because Kant is concerned with the nature of moral action, not its consequences, he isn't concerned with whether autonomous agents achieve better outcomes than heteronomous agents. And yet, the question about the expected outcomes of the different types of agency is an interesting one to pursue, for it is not obvious up front which type of agent would achieve better outcomes. This paper uses game theory to explore and begin to answer this question. We present a game theoretic examination of five forms of heteronomy and their corresponding forms of autonomy. We show that across a significant range of interactive situations agents who have the autonomy to choose between selfishness and either empathy or altruism achieve expected material payoffs equal to the maximum expected material payoffs of the corresponding heteronomous agents. We also show that across the same range of interactions agents who have the autonomy to choose between a deontological moral law and selfishness, empathy, or altruism achieve higher expected material payoffs than the corresponding heteronomous agents.

1 Introduction

Kant distinguishes between autonomous and heteronomous agents. Because Kant is concerned with the nature of moral action, not its consequences, he isn't concerned with whether autonomous agents achieve better outcomes than heteronomous agents. And yet, the question about the expected outcomes of the different types of agency is an interesting one to pursue, for it is not obvious up front which type of agent

*Davidson College. shgourisuresh@davidson.edu
†Davidson College. pastudtmann@davidson.edu

would achieve better outcomes. This paper uses game theory to explore and begin to answer this question.

Traditionally, game theory assumes moral preferences are predetermined (exogenous). Agents strategize based on these external morals, reflecting Kant's heteronomous agents. The standard way of introducing this sort of preference structure is through exogenously defined parameters, and several such parameters have been introduced. For instance, Fehr et al. (2007) introduce a two-parameter utility function that captures inequity aversion, Sally (2001) introduces a three parameter utility function that captures sympathy, Levine (1998) introduces a two-parameter utility function for altruistic and spiteful preferences, and Alger and Weibull (2013) introduce a Kantian preference parameter. Gouri Suresh and Studtmann (2023), however, have proposed a model for morally autonomous agents that does not require parameters. In their models, agents not only make strategic choices but also choose their moral framework. For example, autonomy of the sort Kant described would involve choosing between a selfish strategy (standard game theory's Nash equilibrium) and a deontological strategy based on the categorical imperative. (Studtmann and Gouri Suresh (2021) explore this model in the context of the Prisoner's Dilemma.) This model captures one version of autonomy, but other possibilities exist. First, one might suppose that the moral law is not deontological in nature but rather consequentialist. An autonomous agent might thereby be required to choose between a selfish strategy and a strategy based on maximizing consequences. Second, even if one accepts that the moral law is deontological, there may be versions of autonomy that require the choice between a deontological law and some other moral law, for instance a law of empathy or a consequentialist law. In this paper, we examine five different forms of heteronomy and their corresponding forms of autonomy.

We analyze how these different forms of heteronomy and autonomy perform across all possible symmetric, simultaneous, perfect information, dyadic games. By considering this range of games, our study encompasses three of the most studied games relevant to morality: Stag Hunt, Prisoner's Dilemma, and Hawk-Dove. The importance of these games for morality is well-established. Curry et al. (2019) argue that these games, along with the Nash Bargaining game, form the basis for moral systems across cultures. Similarly, Harms and Skyrms (2008) argue that understanding the evolution of morality requires explaining cooperation in Prisoner's Dilemma, playing stag in Stag Hunt, and playing equal splits in the symmetric bargaining game. Our study is thus an attempt to find a general rule that achieves optimal consequences

across all dyadic, symmetric games rather than specific rules tailored to specific types of symmetric interactions. Although we analyze interactions across a range of games, it's important to note that we only consider games where both players follow the same rule. This specific scenario is necessary to identify which rules lead to the best outcomes assuming everyone follows them. Parfit (2011) calls such rules "optimific" and believes they represent the ideal moral code. While we don't take a position on whether Parfit is right, our analysis can be seen as part of an ongoing search for optimific rules.

In what follows, we begin by examining a parameter that captures a mix of *altruistic* preferences and selfishness. The altruism involved can be seen as inherent in average consequence utilitarianism. We then discuss a parameter which captures a mix of *empathetic* preferences and selfishness. Next, we investigate the Alger and Weibull (2013) Kantian preference parameter, which allows for combinations of *deontological* preferences and selfishness. Finally, we explore parametric combinations of deontology and empathy and of deontology and altruism. We then move on to examine the corresponding forms of autonomy. First, we consider the autonomy to choose between altruism and selfishness. Second, we consider the autonomy to choose between empathy and selfishness. Third, we consider the autonomy to choose between deontology and selfishness. Fourth, we examine the autonomy to choose between deontology and empathy. Finally, we examine the autonomy to choose between deontology and altruism.

By examining these different forms of heteronomy and autonomy, we show the following. In the two cases of autonomy that do not involve deontology, i.e, the autonomy to choose between altruism and selfishness and the autonomy to choose between empathy and selfishness, autonomy achieves the same expected payoffs as the maximum payoff achieved by the corresponding form of heteronomy. One might view such a result as vindicating the power of autonomy. After all, matching the maximum expected value of the corresponding form of heteronomy would seem to be an improvement over heteronomy. However, one might also view the result as showing that autonomy is superfluous. If heteronomy can achieve the same result as autonomy, it is not clear that autonomy does anything that heteronomy cannot. However, in the cases of autonomy that involve deontology, autonomy achieves higher expected payoffs than the maximum expected payoffs of the corresponding forms of heteronomy. These findings raise the broader question of whether deontological autonomy always outperforms its heteronomous counterpart. This paper leaves this question open for

further exploration.

Beyond demonstrating the advantage of deontological autonomy in our investigated cases, our analysis also establishes an absolute benchmark for performance, which is the maximum achievable total payoff across the interactions we examine. The existence of this benchmark allows us to evaluate various forms of autonomy and heteronomy. Our results show that heteronomous agents that we examine achieve expected material payoffs that are between approximately 71% and 81% of the maximum possible payoff, whereas the deontologically autonomous agents that we examine achieve material payoffs between approximately 86% and 96% of the maximum possible payoff.

These results have significant implications for how game theorists should incorporate morality into game theory. Traditionally, game theorists have relied on parameters to represent moral preferences. While these models provide valuable insights, they only depict heteronomous agents with predetermined preferences. This, we contend, misses the essence of moral agency, which requires the ability to choose between different version of the moral law. Of course, some game theorists might not be concerned with capturing the essence of moral agency. They might believe all agents are heteronomous, making parameters a suitable approach. However, our findings reveal a crucial advantage to modeling moral agents as autonomous: deontologically autonomous agents achieve better expected outcomes than heteronomous agents. Therefore, for game theorists interested in modeling agents that achieve high expected outcomes, the goal should be to model agents who choose between deontology and different moral frameworks. This shift has the potential not only to create more nuanced and powerful models of moral decision-making in game theory – indeed, if we are correct it allows for a mathematical investigation into the essence of moral agency – but it should also provide considerable insight into the best way for agents to maximize expected outcomes.

Before diving into the details, we offer a preliminary note on the mathematical findings presented in this paper. In the body of the paper, we present and briefly discuss the utility functions we use to model heteronomous agents as well as the matrices we employ for autonomous agents, and then provide visual plots that vividly illustrate the expected outcomes of these agents' interactions. For readers seeking a quick grasp of our argument, focusing on the plots alongside the concluding table will suffice, as they effectively highlight our conclusions. Additionally, in the appendix, we furnish the mathematical underpinnings behind these results, supplemented by

	s_1	s_2
s_1	A, A	B, C
s_2	C, B	D, D

Table 1. Symmetrical Interaction – Material Payoffs

	s_1	s_2
s_1	A, A	$(B + \alpha C)/(1 + \alpha)$,$(C + \alpha B)/(1 + \alpha)$
s_2	$(C + \alpha B)/(1 + \alpha)$,$(B + \alpha C)/(1 + \alpha)$	D, D

Table 2. Symmetrical Interaction with Altruistic Utility Function

Mathematica worksheets detailing the computations responsible for generating the plots and values in the final table.

2 Heteronomy

2.1 Parametric Combination of Altruism and Selfishness

To incorporate altruism into a game, we can add a fraction of the other player's payoff to a player's own payoff. This fraction is represented by the parameter, α. We suppose that α multiplies the other player's payoff, but to avoid inflated values, we scale everything down by dividing by $(1 + \alpha)$. Let $U_1(x, y)$ be player 1's material payoff and $U_2(x, y)$ be player 2's material payoff. This gives us the following utility function:

$$V(x, y) = (U_1(x, y) + \alpha U_2(x, y))/(1 + \alpha) \tag{1}$$

(This utility function can be derived from the utility function in Levine (1998) by setting its altruism parameter, α, to 1 and its spite parameter, λ, to 0.) For our analysis, We consider the average material payoff of an interaction in which both agents act according to this utility function, assuming the values of the variables in the interaction are uniformly distributed between -1 and 1. We assume that the interaction has a symmetrical set of payoffs as shown in table 1, where s_1 and s_2 are the two strategies available to the agents.The matrix in table 1 contains the objective, or what we call the 'material', payoffs for the interactions we study. The agents in the interaction, however, act as if they are playing a game that has been modified by the above utility function. Hence, they play the game in table 2.

It is important to stress that it is the material (objective) payoffs whose expected value we compute, not the utility of those material payoffs for the agents. The plot

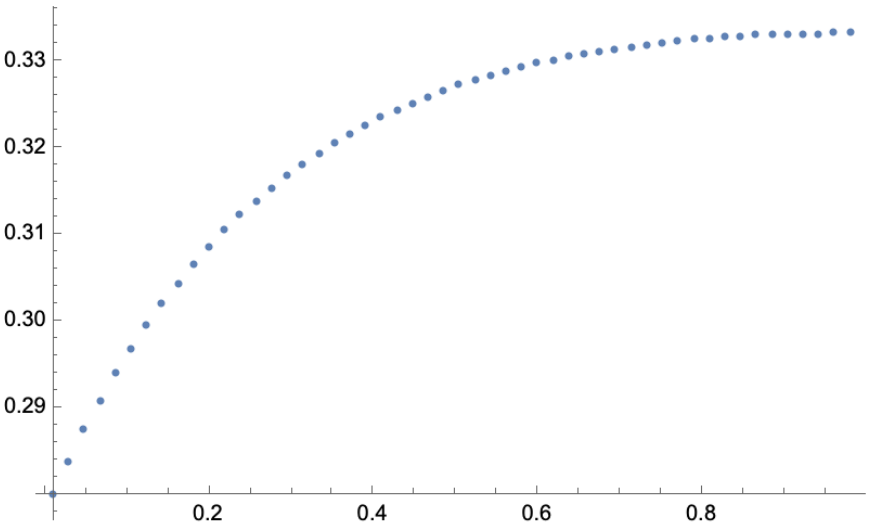

Figure 1. Parametric Combination of Altruism and Selfishness

in figure 1 shows the expected payoff for different levels of altruism with α ranging from 0 to 1. When there's no altruism, $\alpha = 0$, the expected value for both players interacting is at its lowest. As the level of altruism increases, the expected payoff steadily rises with the maximum payoff at $\alpha = 1$. Hence, players who value the well-being of others have greater expected material outcomes than those who do not.

2.2 Parametric Combination of Empathy and Selfishness

There's another way to view altruism: as a trade-off between caring about yourself and others. Let ϵ be a parameter that represents an agent's level of empathy. A higher value for ϵ represents greater empathy, meaning the agent prioritizes the other player's well-being more. When ϵ reaches 1, the agent becomes a complete empath, solely concerned with the other agent's outcome. Once again, let $U_1(x, y)$ be player 1's material payoff and $U_2(x, y)$ be player 2's material payoff. The utility function for this type of altruism is:

$$V(x, y) = ((1 - \epsilon)U_1(x, y) + \epsilon U_2(x, y)) \tag{2}$$

In a two-player game, if the parameter α from the previous section is set to 1 (complete altruism) and ϵ is set to 1/2 (balanced concern), the two utility functions become identical. If we use the same variable to represent both parameters, one can compare the two utility functions. The plot in figure 2 shows the expected payoffs for both parameters along with a line representing the highest expected payoff for both. The

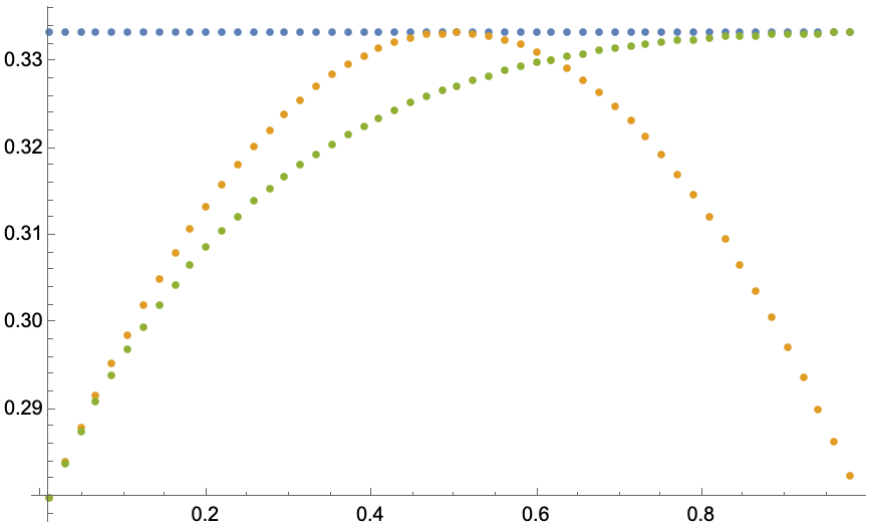

Figure 2. Parametric Combination of Empathy and Selfishness

plots show that both methods of introducing altruism achieve the same maximum outcome. It also shows that empathy achieves the highest expected payoff when it is perfectly balanced between oneself and others, ($\epsilon = 1/2$). Moreover, one can see from the plot for the empathy parameter that maximal empathy, ($\epsilon = 1$), leads to the same expected payoffs as pure selfishness, ($\epsilon = 0$). Both are the lowest possible.

At first glance, it might seem counterintuitive that pure empathy and pure selfishness would lead to the same results. However, this can be understood by considering the nature of the interaction. In the scenarios we examine, both agents are operating under the same principle. When both agents are solely focused on their own benefit, the outcome is identical to when both agents are solely focused on the other's benefit. In the latter case, the other agent's self-interest ends up serving the first agent just as effectively as the first agent's self-interest in the former case. We shall see this phenomenon again when we examine other utility functions.

2.3 Parametric Combination of Deontology and Selfishness

Kantian Deontology can be introduced into game theory via Kantian preferences. These preferences favor situations where everyone behaves similarly to the agent herself. This aligns with the Kantian principle that moral actions should be universalizable. Alger and Weibull (2013) define the Kantian parameter with the following utility function, where κ represents the degree of preference an agent has for outcomes

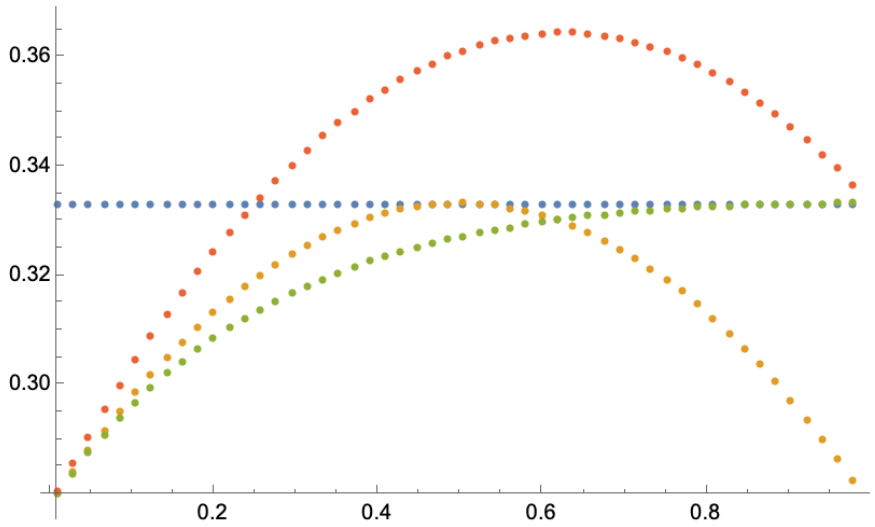

Figure 3. Parametric Combination of Deontology and Selfishness/Empathy

in which everyone acts as she does:

$$V(x, y) = ((1 - \kappa)U_1(x, y) + \kappa U_1(x, x)) \tag{3}$$

If we use the same variable for all the parameters in the three utility functions that we have introduced, we can compare all three.

The plots in figure 3 show the expected outcomes for two agents acting according to the parameterized Kantian utility function as well as the two previously discussed plots of the expected outcomes for two agents acting according to the two different altruism parameters. The plots reveal some interesting facts about Kantian and altruistic preferences. First, Kantian preferences lead to better outcomes than altruistic preferences. Second, the maximum payoff for Kantian preferences occurs around $\kappa = 0.62$, not at $\kappa = 1$. Hence, for agents with Kantian preferences, incorporating some level of self-interest leads to better outcomes.

2.4 Parametric Combination of Deontology and Empathy

We saw in the case of the empathy parameter that pure selfishness and pure empathy achieve the same expected outcomes. This naturally raises the question whether a parametric combination of empathy and deontological preferences would achieve the same expected outcomes as a parametric combination of selfishness and deontological preferences. One can modify the Alger/Weibull Kantian preference parameter so as

to define a parametric combination of empathy and deontology as follows:

$$V(x, y) = ((1 - \kappa)U_2(x, y) + \kappa U_1(x, x)) \tag{4}$$

As it turns out, the material payoffs for two agents acting according to a mix of selfishness and deontology are the same as the material payoffs for agents acting according to a mix of empathy and deontology. In this way, the parametric combinations of selfishness or empathy and deontology are analogous to pure selfishness and pure empathy by themselves. Hence, figure 3 shows the expected outcomes for both types of agent.

2.5 Parametric Combination of Deontology and Altruism

We now turn to the last form of heteronomy we examine. Traditionally, philosophers have viewed consequentialism and deontology as competing theories. The idea of merging these two theories has thus not been explored much. Within game theory, however, a parametric synthesis of the two theories is not hard to introduce. One can define a parameter that reflects an agent's preference between achieving altruistically good outcomes (altruism) and acting according to universalizable moral principles (deontology). If we use a parameter, β, that represents an agent's preference for Kantian outcomes as opposed to altruistic outcomes, the following utility function defines a combined parameter:

$$V(x, y) = (1 - \beta)(U_1(x, y) + U_2(x, y))/2 + \beta U_1(x, x) \tag{5}$$

If we again use a single variable for all the parameters so far introduced, we can plot all the expected material outcomes for all the utility functions so far discussed. Figure 4 shows these expected outcomes. The plots show that a parametric combination of consequentialist and deontological preferences can achieve higher expected payoffs than either approach alone. Interestingly, the optimal value for this parameter occurs when $\beta = 1/2$. This finding is particularly relevant considering the long-standing philosophical debate between deontologists and consequentialists, often viewed as at an impasse. The fact that a balanced blend of these two seemingly opposing viewpoints leads to the best outcomes suggests there might be room for a more nuanced approach to morality than what has heretofore been the common approach of defending one of the views against the other as the correct one.

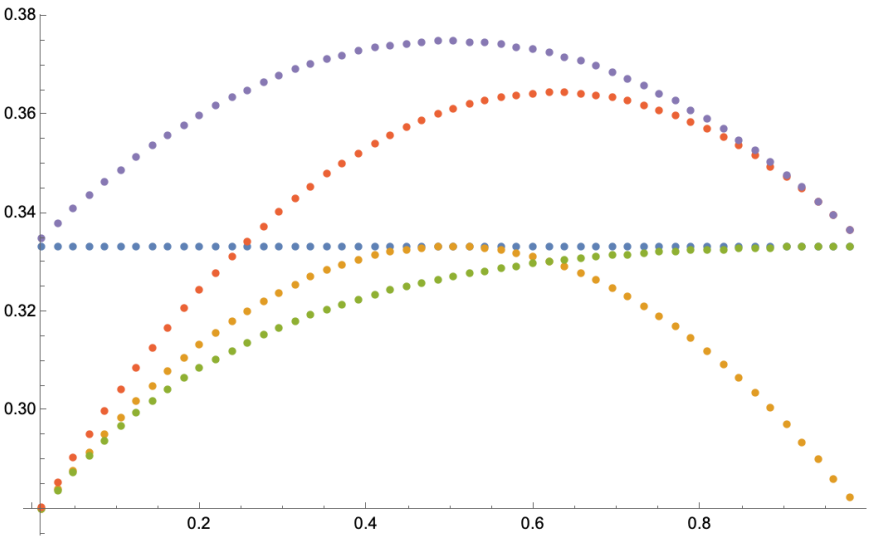

Figure 4. Parametric Combination of Deontology and Altruism

3 Autonomy

After examining five types of heteronomous agents, we now shift our focus to autonomous agents. Kant argued, and we agree, that moral autonomy requires more than simply making a strategic choice given predetermined preferences. Rather, morally autonomous agents must choose the law under which their strategic choices are made. According to Kant, morally autonomous agents must choose between the law of selfishness and a law derived from the categorical imperative. Gouri-Suresh and Studtmann (2024) show that it is possible to define a general mathematical structure that models this type of autonomous choice. The autonomy they define, however, is more general than the autonomy Kant envisions. In their models, agents have not just a strategy space, $S : s_1, ...s_n$, but also a morality space, $M : m_1, ..., m_n$. The expanded strategy space for an autonomous agent, then, is the Cartesian product SxM.

The initial strategy space, S, comes from an underlying game. In the present case, it comes from the symmetrical game in table 1. This raises the question as to where the morality space comes from. Without providing a general answer to this question, we can in the present case generate the morality spaces from the parameters introduced to model heteornomous agents. More specifically, we derive the morality space by extremizing, setting to 0 or 1, the various parameters. This yields two utility functions, i.e., moralities, that an agent can choose in addition to choosing a strategy from the strategy space. In this way, each parameter not only defines a heteronomous

	As_1	As_2	Ss_1	Ss_2
As_1	A, A	$(B+C)/2$	A, A	$(B+C)/2, C$
As_2	$(B+C)/2, A$	D, D	$(B+C)/2, B$	D, D
Ss_1	A, A	$B, (B+C)/2$	A, A	B, C
Ss_2	$C, (B+C)/2$	D, D	C, B	D, D

Table 3. Autonomy to Choose Between Altruism and Selfishness

agent but also defines an associated autonomous agent.

3.1 Autonomy to Choose Between Altruism and Selfishness

The first autonomous agent corresponds to the parametric mix of selfish and consequentialist preferences. Such an autonomous agent chooses between two utility functions (moralities) in addition to making a strategic choice. One utility function represents an agent with the strongest altruistic consequentialist preferences, $\alpha = 1$, while the other represents a purely selfish agent, $\alpha = 0$. The matrix in table 3 represents the interaction between two such autonomous agents, where A represents the decision to be altruistic and S represents the decision to be selfish.

Because this matrix doesn't involve any parameters, we can calculate a single, universal value for the expected material payoffs of agents who play this game. As it turns out, two agents acting as if they are playing the game in table 3 achieve expected material payoffs equal to the expected material payoff when $\alpha = 1$. Hence, this form of autonomy does as well as the corresponding form of heteronomy does at its best.

3.2 Autonomy to Choose Between Empathy and Selfishness

The second autonomous agent corresponds to the empathy parameter, ϵ. Like the previous autonomous agent, this type of autonomous agent chooses between two utility functions that lie at the extremes of the corresponding parameter, namely $\epsilon = 0$ and $\epsilon = 1$. Such agents play the game in table 4, where E represents the decision to be empathetic and S represents the decision to be selfish. Two agents acting as if they are playing the game in table 4 achieve expected material payoffs equal to the expected material payoffs when $\epsilon = 1/2$. Hence, as in the previous case, autonomous agents of this type do as well as the corresponding form of heteronomous agents do at their best.

	Es_1	Es_2	Ss_1	Ss_2
Es_1	A, A	C, B	A, A	C, C
Es_2	B, C	D, D	B, B	D, D
Ss_1	A, A	B, B	A, A	B, C
Ss_2	C, C	D, D	C, B	D, D

Table 4. Autonomy to Choose Between Empathy and Selfishness

	Ds_1	Ds_2	Ss_1	Ss_2
Ds_1	A, A	A, D	A, A	A, C
Ds_2	D, A	D, D	D, B	D, D
Ss_1	A, A	B, D	A, A	B, C
Ss_2	C, A	D, D	C, B	D, D

Table 5. Autonomy to Choose Between Deontology and Selfishness

3.3 Autonomy to Choose Between Deontology and Selfishness

The third autonomous agent corresponds to the Kantian preference parameter from Alger and Weibull. While a heteronomous agent with this preference has a preset mix of self-interest and deontological concern, the autonomous agent chooses between two utility functions in addition to making a standard strategic choice. One utility function represents an agent with the strongest Kantian preference, ($\kappa = 1$), while the other represents a purely selfish agent,($\kappa = 0$). The matrix in table 5 represents the interaction between two such autonomous agents, where D represents the decision to be deontological and S represents the decision to be selfish.

Unlike the previous two cases, the expected outcome of agents playing the game in 5 achieve higher expected material payoffs than the maximum possible for the corresponding form of heteronomy. The graph in figure 5 shows the expected payoff for this form of autonomy along with the previous plots for a comparison. The plot reveals a fascinating outcome: autonomous agents who can choose between following a Kantian moral principle and acting out of self-interest (selfishness) achieve better expected payoffs than agents with predetermined preferences (heteronomous agents). In fact, this type of autonomy surpasses all the forms of heteronomy we've explored so far. This strongly suggests that the ability to make deontologically autonomous moral choices is a key to achieving better outcomes.

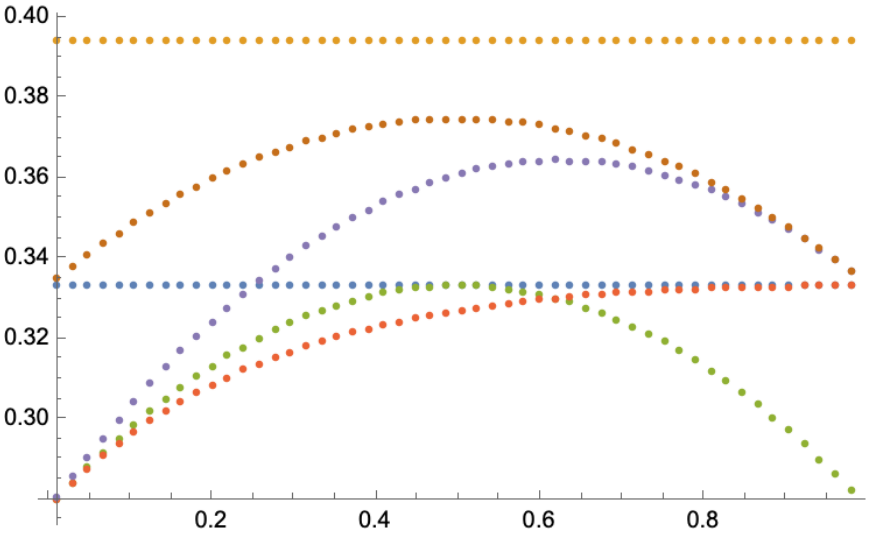

Figure 5. Autonomy to Choose Between Deontology and Selfishness/Empathy

	Ds_1	Ds_2	Es_1	Es_2
Ds_1	A, A	A, B	A, A	A, B
Ds_2	D, A	D, B	D, C	D, D
Es_1	A, A	C, B	A, A	C, B
Es_2	B, A	D, B	B, C	D, D

Table 6. Autonomy to Choose Between Deontology and Empathy

3.4 Autonomy to Choose Between Deontology and Empathy

As we have discussed, the parametric combination of deontology and selfishness achieves the same expected outcomes as the parametric combination of deontology and empathy. This naturally raises the question whether the autonomy to choose between deontology and selfishness achieves the same expected outcome as the autonomy to choose between deontology and empathy. The matrix in table 6 represents two agents who have the capacity to choose between Kantian deontology and empathy, where D represents the decision to be deontological and E represents the decision to be empathetic.

Perhaps not suprisingly, the expected payoff for this form of autonomy is equal to the expected payoff for the previous form of autonomy. Hence, this form of autonomy outperforms the corresponding form of heteronomy.

	Ds_1	Ds_2	As_1	As_2
Ds_1	A,A	A,D	A,A	A,$(B+C)/2$
Ds_2	D,A	D,D	D,$(B+C)/2$	D,D
As_1	A,A	$(B+C)/2$,D	A,A	$(B+C)/2$,$(B+C)/2$
As_2	$(B+C)/2$,A	D,D	$(B+C)/2$,$(B+C)/2$	D,D

Table 7. Autonomy to Choose Between Deontology and Altruism

3.5 Autonomy to Choose Between Deontology and Altruism

We previously saw that a parameterized combination of altruistic and deontological preferences led to better outcomes than the parametric combination of deontological preferences and selfish preferences. This raises the question: would the autonomy to choose between altruism and deontology outperform the autonomy to choose between selfishness and deontology? The matrix in table 7 represents the interaction between two autonomous agents who can choose between acting on altruistic principles or deontological principles, where D represents the decision to be deontological and A represents the decision to be altruistic.

Similar to the previous matrices for autonomous agents, this matrix is parameter-free, which allows us to calculate a single, universal value for the expected material payoffs of agents following this strategy. The graph in figure 6 shows the expected outcome for the choice between deontology and altruism along with the previous plots for a comparison of all the approaches so far discussed. The plots reveal that agents who can choose between altruism and deontological principles outperform all the previous agents with predetermined preferences (heteronomy) as well as agents with the autonomy to choose between deontology and selfishness. Hence, autonomous agents of this type outperform their heteronomous counterparts. Moreover, the autonomy to choose between deontology and altruism outperforms the autonomy to choose between deontology and selfishness. In fact, this type of autonomy comes very close to achieving the absolute best possible outcome.

4 The Maximum Total Expected Payoff

The near optimality of the autonomy to choose between deontology and altruism can be seen by considering the original symmetric form in table 1. There are three possible values for the total expected outcome in that game: $2R$, $2P$, $T+S$. Hence, in terms of total expected value, there are six possible orderings. Because the orderings for the

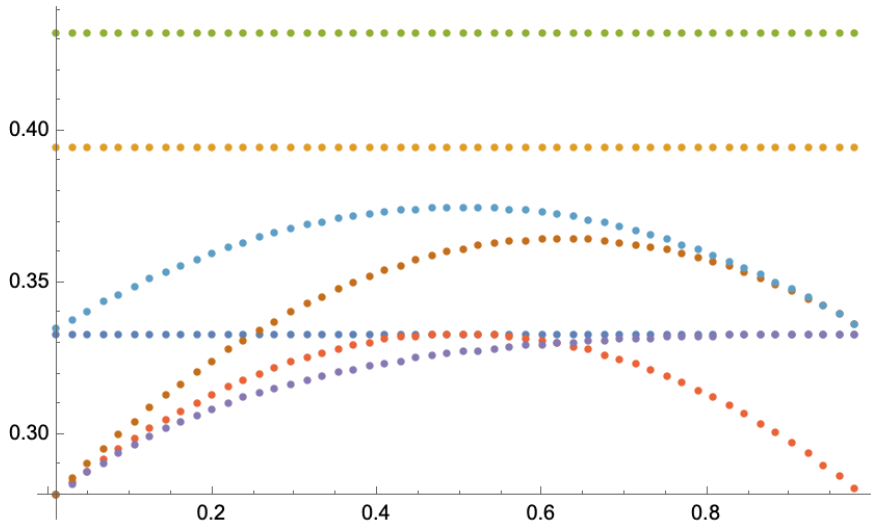

Figure 6. Autonomy to Choose Between Deontology and Altruism

conditions in which $R > P$ are symmetrical to the conditions in which $P > R$, we need consider only three possible orderings: (1)$R > P > (T+S)/2$; (2)$R > (T+S)/2 > P$; and (3)$(T + S)/2 > R > P$. When (1) or (2) is the case, the best possible outcome occurs when both agents coordinate on the action that has R as an outcome. When (3) is the case, the best possible outcomes occur when the two agents anti-coordinate. Now, when one considers the matrix for the form of autonomy we are considering, it is evident that the individual players' payoffs are equal to R, P, or $(T+S)/2$. Hence, the solution space for the game has the same ordering as the total expected outcomes in the original symmetric game. When (1) or(2) is the case both agents coordinate so as to get the payoff equal to R, which is the way to achieve the maximal total outcome under those conditions. When (3) is the case, the game becomes an anti-coordination game with three equilibria: two pure strategy asymmetric equilibria and one symmetric mixed strategy equilibria. To the extent that the agents play the asymmetric equilibria, they achieve the maximum possible total payoff. The only thing keeping them from always achieving the highest total expected payoff, therefore, is the symmetric mixed-strategy equilibrium.

This suggests, then, a restriction: restrict the equilibria that the agents choose to pure-strategy equilibria. Call such a moral rule 'Pure Strategy Deontarianism.' Pure Strategy Deontarianism is an artificial rule, since it involves a mathematically artificial restriction. Nonetheless, it does entail that the agents in question achieve the maximum possible expected payoffs. And that is a useful fact. For it is possible to compute the expected payoffs for agents who act according to Pure Strategy

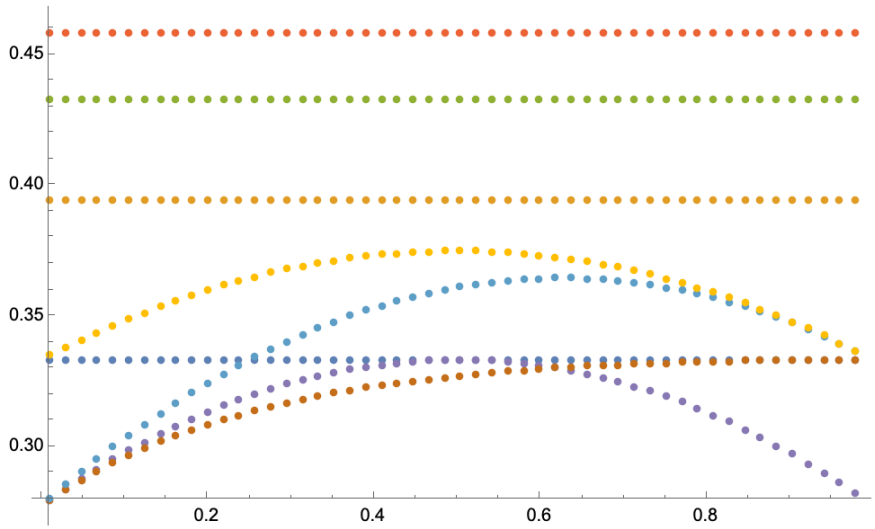

Figure 7. Maximum Total Expected Payoffs

Moralities	Heteronomy	Autonomy
Altruism and Selfishness	71%	71%
Empathy and Selfishness	71%	71%
Deontology and Selfishness	79%	86%
Deontology and Empathy	79%	86%
Deontology and Altruism	81%	96%

Table 8. Heteronomy Versus Autonomy

Deontarianism, which, because it is a maximum, gives an absolute measure against which to judge the outcomes of the other moral rules. The plot in figure 7 shows the maximum possible total expected payoffs along with all the previous plots.

We can now calculate how close the different moral rules come to achieving the absolute best possible outcome. Table 8 shows the percentages of the maximum total possible outcome for all the various heteronomous and autonomous agents we have examined.

These numbers highlight a key finding: autonomous agents who can strategically choose between a deontological law and some other consequence-based law, whether it be selfishness, empathy, or altruism, outperform corresponding heteronomous agents with a predetermined combination of preferences. And independent of a comparison to heteronomy, the results for autonomous agents are impressive: choosing between deontology and selfishness or deontology and empathy achieves 86% of the maximum possible outcome; and choosing between deontology and altruism gets particularly close at 96%.

5 Concluding Remarks

After presenting our mathematical analysis, we conclude this paper by discussing very briefly some philosophical implications of our findings. The most immediate implication is that in the scenarios we examined, autonomous deontology proves superior to heteronomy. This outcome can be seen as a validation of Kant's distinction, although it relies on expected outcomes, a concept foreign to Kant's original framework. It's important to note the limitations of our study. Firstly, we focused on just three forms of deontological autonomy, suggesting that an autonomous synthesis of deontology and other moral preferences generally outperforms heteronomous synthesis. However, a broader investigation would be needed to verify this claim mathematically. Additionally, our study only considered cases where both agents adhere to the same rule. Future research should extend to interactions involving moral agents and selfish agents. Lastly, our analysis was limited to symmetrical interactions. Nevertheless, even within these constraints, the superiority of autonomy over heteronomy across such an important range of interactions amply demonstrates the potential advantages of autonomous agency.

Hegel's famous idea that history progresses through a dialectical process involving thesis, antithesis, and synthesis is widely known and has profoundly influenced theoretical discourse. Until now, however, there has been no clear demonstration, to our knowledge, that such a theoretical synthesis constitutes a genuine improvement within any theoretical domain. Kantian deontology and utilitarianism, arguably representing opposing theories akin to thesis and antithesis, provide a compelling case study. Our findings unequivocally demonstrate that across the range of interactions we examine agents, whether autonomous or heteronomous, who act according to a synthesis of these two moralities (moral preferences) achieve higher expected outcomes than agents who act according to either separately. Hence, one of the key philosophical insights from our study is the unexpected validation of Hegel's perspective within the moral domain—a truth that eluded even Hegel himself, who critiqued both Kantian deontology and utilitarianism separately without suggesting their synthesis.

Another significant philosophical implication of our results is that the autonomous synthesis of deontology and utilitarianism approaches the maximum expected total outcome, which validates the central roles these theories play in moral philosophy. While moral philosophers may not have previously considered combining these theories, they deserve recognition for articulating frameworks whose synthesis approaches optimality. However, the gap between the expected value of this synthesis and the

theoretical maximum suggests that deontology and utilitarianism alone do not fully explain moral decision making. There are likely other rules that achieve outcomes within this gap, suggesting avenues beyond the standard models of moral philosophy. Exploring morality beyond these standard models promises profound insights into optimal moral decision making processes.

References

Alger, Ingela and Jörgen W. Weibull, "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching," *Econometrica*, 2013, *81* (6), 2269–2302. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA10637.

Curry, Oliver, Daniel Austin Mullins, and Harvey Whitehouse, "Is It Good to Cooperate: Testing the Theory of Morality-as-Cooperation in 60 Societies," *Current Anthropology*, 2019, *60* (1), 47–69.

Fehr, Ernst, Alexander Klein, and Klaus M Schmidt, "Fairness and Contract Design," *Econometrica*, 2007, *75* (1), 121–154. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2007.00734.x.

Harms, William and Brian Skyrms, "Evolution of Morality," in "Oxford Handbook of Philosophy of Biology," Oxford New York: Oxford Univeristy Press, 2008, pp. 434–450.

Levine, David K., "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, July 1998, *1* (3), 593–622.

Parfit, Derek, *On What Matters, Volumes 1 and 2*, Oxford University Press, 2011.

Sally, David, "On sympathy and games," *Journal of Economic Behavior & Organization*, January 2001, *44* (1), 1–30.

Studtmann, Paul and Shyam Gouri Suresh, "Universalizing and the We: Endogenous Game Theoretic Deontology," *Economics and Philosophy*, 2021, *37* (2), 244–259. Publisher: Cambridge University Press.

Suresh, Shyam Gouri and Paul Studtmann, "Angels and Devils on Our Shoulders: A Framework for Modeling Moral Agency," January 2023.