


ARTICLE

Universalizing and the we: endogenous game theoretic deontology

Paul Studtmann¹ and Shyam Gouri Suresh^{2,*} 

¹209 Ridge Road, Department of Philosophy, Davidson College, Davidson NC 28036, USA and ²209 Ridge Road, Department of Economics, Davidson College, Davidson NC 28036, USA

*Corresponding author. Email: shgourisuresh@davidson.edu

(Received 21 August 2019; revised 2 July 2020; accepted 17 July 2020; first published online 25 September 2020)

Abstract

The Nash counterfactual considers the question: what would happen were I to change my behaviour assuming no one else does. By contrast, the Kantian counterfactual considers the question: what would happen were everyone to deviate from some behaviour. We present a model that endogenizes the decision to engage in this type of Kantian reasoning. Autonomous agents using this moral framework receive psychic payoffs equivalent to the cooperate-cooperate payoff in Prisoner's Dilemma regardless of the other player's action. Moreover, if both interacting agents play Prisoner's Dilemma using this moral framework, their material outcomes are a Pareto improvement over the Nash equilibrium.

Keywords: Kantian Reasoning; Morality in Game Theory

1. Introduction

Humans have a moral sense. To capture this sense game theorists have recently introduced the Kantian counterfactual into game theory. The Nash counterfactual considers the question: what would happen were I to change my behaviour on the assumption that no one else does. By contrast, the Kantian counterfactual considers the question: what would happen were everyone to deviate from some behaviour? There are currently two extant approaches to the Kantian counterfactual. Both reserve a special place for situations that conform to Kant's categorical imperative, i.e. situations in which everyone acts in the same way. Alger and Weibull (2013, 2016) introduce an exogenous parameter, k , into an agent's utility function that indicates the degree to which the agent considers the Kantian outcome in which everyone acts the same as she does. Roemer (2010, 2015, 2019) introduces a distinct solution concept that requires agents to choose among situations in which everyone acts the same.

The mathematical results of models that include the Kantian counterfactual are impressive and many applications have been suggested. Roemer (2010, 2015) has proven that agents conforming to his solution concept achieve Pareto optimal results in all monotone games. He argues:

© The Author(s), 2020. Published by Cambridge University Press.

While economists schooled in Nash equilibrium may view the Kantian behavior as utopian, there is some – perhaps much – evidence that it exists. If cultures evolve through group selection, the hypothesis that Kantian behavior is more prevalent than we may think is supported by the efficiency results here demonstrated. (Roemer 2015: 45)

Alger and Weibull (2013) have argued that evolution selects a degree of Kantianism, or what they call a degree of morality, that equals the assortativity index, which is a key parameter in population statistical analysis. In their words, “It is as if homo moralis with degree of morality equal to the index of assortativity preempts mutants; any rare mutant can at best match the payoff of the residents” (2013: 2272). They go on to argue that their models have applications to interactions between kin, business partners and geographically close agents. Newton (2017) has shown how conditions faced by primitive societies could have led to the evolution of the ability to collaboratively share intentions in the way that Kantian models entail. And Grafton *et al.* (2017) argue that Kantian models better explain the international response to global warming than do Nash models.

It is the purpose of this paper to discuss a third type of Kantian model, one that introduces Kantian reasoning by way of a recursively defined action type that we call ‘universalizing’. Although our model shares with other Kantian models the central idea that Kantian reasoning involves an agent’s considering what would happen were everyone to play the same strategy that she is playing, it differs in that it makes an agent’s moral behaviour depend on a decision based upon endogenous features of the world rather than on something that is exogenously or evolutionarily determined. There are some advantages to approaching Kantian reasoning in this way.

Modelling human interactions by way of the Kantian counterfactual without allowing agents to choose what level of Kantian behaviour to exhibit deprives them of a core feature of moral agency. Allowing agents to choose the extent to which they universalize their actions, on the other hand, gives them a choice that is essential to autonomy. While moral preferences or moral behaviour may have emerged evolutionarily, a question remains as to how an agent, one who can choose her level of moral behaviour, would behave. Although the evolutionary genesis of moral behaviour may be the result of non-autonomous agents responding to a multitude of exogenous factors, one might still want to know how agents, were they ever to have the capacity to choose their level of moral engagement, would behave.

By including universalizing as a distinct recursively defined action type, our model provides the resources to analyse such agents. The recursive definition makes the models solvable by way of the standard Nash solution concept. As a result, our approach allows for standard mixed strategy Nash equilibria over action types where one of the types is the decision whether to universalize or not. Hence, our model does not require any appeal to an exogenous factor like the assortativity index, as is found in Alger and Weibull (2013); it does not require an appeal to the evolution of specific moral preferences as described in Roemer (2015) and Newton (2017), in order to explain Kantian behaviour; nor does it require a solution concept distinct from the standard Nash solution

concept as is found in Roemer (2015) and Grafton *et al.* (2017). The recursive nature of the model we propose demonstrates that Kantian models can be integrated into the Nash universe with Nashian reasoning as the basis for both rationality and surprisingly enough morality as well.

In contributing to an analysis of autonomy, the recursive nature of our model allows for a mathematical analysis of various aspects of what might be considered the noumenal and phenomenal realms. We use the terms ‘phenomenal’ and ‘noumenal’ with some reservation, since our understanding of them will deviate from Kant’s own understanding. We consider a noumenal agent to be an autonomous individual who is capable of choosing his level of moral engagement with his interactive partner by way of endogenous features of a situation and a phenomenal agent to be an individual whose choices are determined by exogenous features inherent in evolutionary processes. Although Kant denies that we can know the noumenal realm, we shall argue that our models provide some insight into the nature of noumenal agents. One might consider this a reason not to use the terms that Kant introduced. But we find the distinction between two different aspects of the world to be rather naturally suggested by the mathematical results that we discuss in the last section of the paper and so, reservations aside, employ the Kantian terminology for our own ends. We stress that nothing substantive hangs on our decision to employ the Kantian terminology.

The first aspect of the noumenal that our model makes both vivid and precise is that autonomy requires what can be called *psychic payoffs* as distinct from *material payoffs*. The appeal to psychic payoffs or utility functions based on more than just an individual’s own material payoffs is common in game theory as a way of explaining the predictive failure of various games. Some such appeal has been made by Sen (1974, 1977), Rabin (1993), Binmore (1994), Bolton and Ockenfels (2000), Falk and Fischbacher (2006) among others. The main difference between our approach and other approaches is that our psychic payoffs are entailed by the decision to universalize. As shall become apparent in the model we examine, universalizing entails psychic payoffs that have a motivational structure analogous to the *moral emotions* of guilt and forgiveness. Our model thus provides an analytic link between autonomy and moral emotions. It may be worthwhile to emphasize that our approach to these moral emotions is through a philosophical framework rather than through an empirically observed analysis of human behaviour.

The inclusion of psychic payoffs that can differ from material payoffs has another consequence: it allows for an analysis of we-interactions. As shall become clear, the label ‘universalizing’ for our purposes is synonymous with ‘playing a we-strategy’. When an agent in an interaction universalizes, she plays a we-strategy and in so doing is capable of experiencing psychic payoffs that are motivationally analogous to guilt and forgiveness. The noumenal, on our account, is thus populated not only with emotions but with the we. Playing a we-strategy connects an autonomous agent to her interactive partner by way of fundamental moral emotions. The model in this paper thus corroborates the arguments of various philosophers – for instance Tuomela and Miller (1988), Searle (1990) and McCann and Bratman (1991) – according to which there are collective acts that are distinct

from individual acts. According to our analysis, such collective acts have an inherently normative basis.

The ability to define the-we in terms of universalizing has implications for the philosophy of collective action. The appeal to *team reasoning* is one of the main approaches in an effort to understand collective action. In Bacharach's version (1997), people we-reason when they bring to a situation a frame that includes the we-concept. He posits a parameter, ω , that gives the probability that a person we-identifies. Bacharach treats ω as exogenous. He was, however, rightly dissatisfied with the exogeneity of ω – to treat the decision to enter into a we-interaction with another agent as exogenous eliminates the fundamental agency involved in being a we – and sought (unsuccessfully) to endogenize it (Smerilli 2014). Because our model allows one to define a we-strategy in terms of universalizing, and because universalizing is an endogenous decision, our model fills in this lacuna in Bacharach's view.

The remainder of this paper is structured as follows. In section 1, we augment a standard Prisoner's Dilemma with the action type of universalizing – we call the resulting model *Universalized Prisoner's Dilemma* (UPD) – and discuss its interpretation with an eye to philosophical issues. We restrict our attention in this paper to UPD, reserving for another paper a more general mathematical treatment of universalized games. Our goal in this paper is thus a modest one. By focusing on the Prisoner's Dilemma, we aim to explore some of the philosophical and mathematical aspects of universalizing as they pertain to a single game, albeit one that has played a central role in discussions of social cooperation. In section 2, we discuss the Nash equilibria of UPD. As shall become apparent, there are three equilibria – two asymmetric pure strategy equilibria and one symmetric mixed strategy equilibrium. We argue on philosophical grounds that the symmetric equilibrium is of primary importance for an autonomous agent. We go on to discuss in detail the behaviour of an agent who acts optimally in the symmetrical equilibrium. Finally, in section 3, we discuss the psychic and material payoffs of an agent who acts optimally in the symmetrical equilibrium and suggest there that the model we have proposed provides an insight into the relationship between the noumenal and phenomenal realms.

2. Universalized Prisoner's Dilemma (UPD)

We begin with the following standard representation of a Prisoner's Dilemma (Table 1).

In this game, if two cooperators, C, interact, each gets the payoff R , the 'reward for mutual cooperation'. If a cooperator meets a defector, D, the cooperators gets S , the 'sucker's payoff', while the defector gets T , the 'temptation of defection'. If two defectors interact, each obtains the payoff P , the 'punishment' of mutual defection. The game is a prisoner's dilemma if $T > R > P > S$. Although there are different possible representations of PD, they can all be transformed into this one by way of positive affine transformations.

To universalize a game requires adding to the initial action type, in the present case cooperate or defect, a second action type: universalize. When a player

Table 1. Prisoner's Dilemma (PD)

	C	D
C	R, R	S, T
D	T, S	P, P

universalizes, he receives as a payoff what he would have received in the original game had everyone played the strategy he is playing. By including universalizing as a distinct action type, we are allowing agents to choose their level of Kantian behaviour based on endogenous features of the model. A mixed strategy in our model is thus a standard Nash mixed-equilibrium; but it is a mixed equilibrium in which an agent chooses to universalize (act as a Kantian), with some probability and chooses not to universalize (act as a Nashian), with some probability.

So, for instance, if a player chooses to cooperate, if he also universalizes, he receives the payoff, R , which is what he would have received in PD were the other player to cooperate as well. If, on the other hand, a player chooses to cooperate without universalizing, he receives a payoff that depends on what the other player does: if the other player cooperates, then the first player receives the reward for cooperating, R ; but if the other player does not cooperate, then the first player receives the sucker's payoff, S . By playing a universalized game, one in effect plays PD with an ability that allows one to play among like-minded agents. Playing a universalized game can thus be seen as a way of forming a 'we-intention'.

Table 2 presents the result of adding universalizing to PD.

Considering this payoff matrix briefly should provide some content to the concept of universalizing, which might seem an odd ability to possess. How, after all, can one act in a way that guarantees that one receives a payoff as if the other person played the very same action? The short answer to this question is: one receives psychic payoffs that can differ from material payoffs.

To see the need for psychic payoffs in understanding the payoff matrix in Table 2, consider the square in the first column of the second row of payoffs, which represents the outcome in which player 1 universalizes and defects and player 2 universalizes and cooperates. In a Prisoner's Dilemma in which player 1 defects and player 2 cooperates, player 1 would receive T , the temptation to defect, and player 2 would receive S , the sucker's payoff. But, because both players universalize, player 1 receives P , the punishment for defecting, while player 2 receives R , the reward for cooperating. As a result of universalizing, player 1 is penalized for defecting – his payoff is less than the payoff he would have received in PD – while player 2 is compensated for cooperating – her payoff is greater than the payoff she would have received in PD. As a result of universalizing, a player's payoff can differ from the material payoffs of a Prisoners Dilemma.

The material payoffs for those who play UPD can be calculated from a payoff matrix that replicates PD in all four two by two quadrants as presented in Table 3.

One can see by inspecting the material payoff matrix in Table 3 and comparing it with the previous table displaying psychic payoffs that the two differ for an agent

Table 2. Universalized Prisoner's Dilemma (UPD): psychic payoffs

	UC	UD	~UC	~UD
UC	R, R	R, P	R, R	R, T
UD	P, R	P, P	P, S	P, P
~UC	R, R	S, P	R, R	S, T
~UD	T, R	P, P	T, S	P, P

Table 3. Universalized Prisoner's Dilemma (UPD): material payoffs

	UC	UD	~UC	~UD
UC	R, R	S, T	R, R	S, T
UD	T, S	P, P	T, S	P, P
~UC	R, R	S, T	R, R	S, T
~UD	T, S	P, P	T, S	P, P

only if she universalizes. Moreover, psychic and material payoffs differ only when a symmetry is broken, i.e. when one agent cooperates and the other defects.

The psychic payoffs involved in asymmetrical we-interaction function like guilt and forgiveness. When an agent universalizes and defects, he receives as a payoff *P*, the punishment for defecting. Although defecting against a cooperator is materially beneficial, when an agent plays universalized defect against a cooperator she experiences a psychic punishment. In this way, universalizing introduces a motivational structure that is analogous to the motivational structure of guilt. When a moral agent acts immorally, his material gain from his action is offset by the negative internal psychic state of guilt. It is easy to see from the payoff matrix that the strategy of universalized cooperation (UC) strictly dominates the strategy of universalized defection (UD). Hence, when a rational agent plays a we-strategy, she will cooperate. This can be seen as the motivational power of guilt.

On the other hand, when an agent universalizes and cooperates, she receives as a payoff, *R*, the reward for cooperating. Her cooperation may be met with cooperation from her interactive partner in which case her psychic payoff, *R*, matches her material payoff. But, her cooperation may also be met with defection from her interactive partner. In such a case, she receives the punishment payoff, *P*, materially but nonetheless receives the reward for cooperation, *R*, psychically. Playing morally against a defector comes with psychic reward. In this way, universalizing introduces a motivational structure analogous to the motivational structure of forgiveness. When a moral agent suffers a harm from someone, his material loss is accompanied by the psychic tranquillity that accompanies forgiveness.

Our theory has three philosophical consequences. First, agents who universalize play a we-strategy. To the extent that playing a we-strategy is equivalent to having a we-intention, being a we according to our theory requires an appropriate sort of

intention. Our theory is thus in broad agreement with several other theories – for example Tuomela and Miller (1988), Searle (1990), McCann and Bratman (1991), Gilbert (1992), Roth (2004), Chant and Ernst (2008) – according to which collective activity requires an intention to act collectively. Second, playing a we-strategy leaves an agent open to psychic payoffs that differ from material payoffs and thereby introduces a moral consideration of one's interactive partner. Finally, and related to this last feature, when you and I interact as a we, we place normative demands on each other, as can be seen by the psychic adjustments due to non-cooperation. Our placement of moral demands on each other, however, is the result of our individual decision to act autonomously as a we. By making the decision to act as a we, I impose on myself a set of motivations that make me morally responsive to you.

A full account of the way we view the nature of morality and its relationship to the-we is beyond this paper. It would be hard, however, to present a better summation of what our theory would look like than Wallace (2013) does, who presents a compelling case for it.

In a situation in which I do something morally wrong, the person adversely affected will have been wronged by me, and have a privileged basis for moral complaint, resentment, and so on, precisely insofar as I have acted with indifference to the value of relating to them on a basis of mutual recognition and regard. The very principles that specify what I have moral reason to do, on this relational conception, equally serve to specify normative expectations and entitlements on the part of others. Those principles are thus implicated in a directed normative nexus very like the one that defines the reciprocal reasons and expectations constitutive of a relationship of friendship. This is a way of thinking about the normative significance of morality that is quite unlike the teleological conception upon which consequentialist approaches rely. (Wallace 2013: 163)

We agree with Wallace that the normative nexus essentially involves a relation to another. Although the other-directedness of a Kantian morality is sometimes obscured by its emphasis on universal maxims, approaching Kantian morality within game theory by way of a recursively defined action type makes explicit that a rational autonomous agent is directed to cooperate with another by way of moral emotions that are entailed by his autonomous decision to act as a we.

3. Autonomous behaviour

In the previous section we introduced universalizing as a recursively defined action type and showed that the resulting model entails psychic payoffs that are motivationally analogous to the moral emotions of guilt and forgiveness. Such a fact can be considered an analytic demonstration of the content of the concept of autonomy. That UPD contains psychic payoffs distinct from material payoffs is entailed by the nature of universalizing, which itself is a choice that autonomous agents, by definition, are capable of making. The articulation of the analytic content

of autonomy, however, falls short of providing a moral imperative. In this way, UPD is in line with Kant who considered his categorical imperative to be a synthetic a priori truth, not an analytic truth. To derive his imperative, Kant appealed to the conditions required for autonomous action. Within a game-theoretic context, on the other hand, one need not appeal to the conditions of autonomous action but can instead appeal to the Nash equilibria of universalized games.

Determining a rule of morality by way of UPD, however, is not a completely straightforward affair. It is not as simple as determining the set of equilibria in UPD and insisting that an agent ought morally to play according to any one of the strategies. For, it may be that an equilibrium does not represent a universalizable strategy in a sense of ‘universalizing’ that is distinct from the sense involved in the action type that appears in UPD. If a rule of behaviour in this second sense of universalizing is universalizable, all agents must universalize in the first sense of ‘universalizing’ with the same probability. If not, then the moral demand on one agent would differ from the moral demand on another agent, which conflicts with an idea central to morality, namely that all agents are equal in the face of morality. Because autonomy involves this second sense of universalizing, deciding the rule a moral agent ought to adopt is not as simple as discovering the Nash equilibria of UPD. For, it may be that such equilibria are asymmetric and require agents to universalize with different probabilities. Instead, an autonomous agent must choose among the *symmetric* equilibria of UPD. Having done so, an autonomous agent can then apply the rule of behaviour to the Prisoner’s Dilemma in which he finds himself.

This extra complexity in the concept of autonomy lends itself to an interpretation of the role of UPD. One might initially suppose that in the context of a Prisoner’s Dilemma an autonomous agent transforms PD into UPD and plays the latter. Such an interpretation, however, does not guarantee that an autonomous agent will land on the correct moral rule. On the other hand, one might suppose that UPD is a structure that autonomous agents use to figure out *how they will play the Prisoner’s Dilemma*. Having used UPD to determine which rules (or rule) are involved in its symmetrical equilibria, an autonomous agent can then apply some such rule of behaviour to the Prisoner’s Dilemma in which he finds himself. This way of viewing autonomy bears some similarity to Roemer’s Kantian optimizing. Whereas a Kantian optimizer restricts the strategy space to symmetrical strategies and then chooses a strategy that maximizes his welfare, our autonomous agent maximizes his welfare on the assumption that he is capable of universalizing and then chooses among the resulting symmetrical equilibrium strategies.

What, then, are the equilibria of UPD? There are three. Two of the equilibria are asymmetric pure-strategy equilibria and one is a symmetric mixed-strategy equilibrium. To see what the equilibria are, it is helpful to note that in UPD the strategy UC strictly dominates UD. Hence, for the purposes of determining the equilibria, the matrix reduces to a three by three matrix (Table 4).

Because $T > R > P > S$, the pure strategies: UC/\tilde{UD} and \tilde{UD}/UC are in equilibrium. These two strategies are precisely the sort of strategies that are ruled out by the condition that an autonomous agent assumes his counterparty to be autonomous as well. That leaves the symmetric mixed-strategy equilibrium as the source for the moral rule governing the Prisoner’s Dilemma.

Table 4. Universalized Prisoner’s Dilemma (UPD) reduced

	UC	UD	~UC	~UD
UC	R, R	$R_1 - P_1$	R, R	R, T
UD	$P_1 - R_1$	$P_1 - P_1$	$P_1 - S_1$	$P_1 - P_1$
~UC	R, R	$S_1 - P_1$	R, R	S, T
~UD	T, R	$P_1 - P_1$	T, S	P, P

The mixed strategy in equilibrium is a mixture of *UC* and \tilde{UD} . The strategies in this mixture already contain an important philosophical result. It may seem obvious that cooperation would be paired with the decision to universalize, since it may seem obvious that cooperating in a Prisoner’s Dilemma is the moral thing to do. Obviousness, aside, however, the pairing of universalizing and cooperation is not analytically contained in the concept of autonomy. Were one to appeal to the conceptual framework that Kant employed, one could consider this basic mathematical result as a demonstration of a synthetic a priori connection between the concepts of universalizing and cooperation.

The probability that an agent plays *UC*, $Pr(UC)$, is given by the following equation:

$$Pr(UC) = (R - P)/(T - P)$$

This equation has a number of consequences. The first and most obvious consequence is that an autonomous agent would not always play a universalized (in the first sense of ‘universalizing’) strategy. Hence, despite the longstanding and exalted position that Kant’s categorical imperative has occupied in philosophical discussions, a game theoretic treatment of autonomy shows that Kant’s imperative is false as long as the concept of universalizing in his imperative is understood in the first, rather than the second, sense of universalizing that we have distinguished. Rather than always playing a universalized strategy an autonomous agent plays a universalized strategy with a probability that depends on three factors: the reward for cooperation, *R*, the punishment for defection, *P*, and the temptation to defect, *T*. The partial derivatives of $Pr(UC)$ with respect to each of these variables provide the basic pattern of behaviour of an autonomous agent.

$$\frac{\partial Pr(UC)}{\partial R} = 1/(T - P)$$

$$\frac{\partial Pr(UC)}{\partial P} = (R - T)/(T - P)^2$$

$$\frac{\partial Pr(UC)}{\partial T} = (P - R)/(T - P)^2$$

Given that $T > R > P$, the first of these derivatives is greater than zero and the second and third of these derivatives are less than zero.

These derivatives describe stakes and temptation dependent cooperative behaviour. In a Prisoner’s Dilemma, the difference between the reward for

cooperation and the punishment for defection, $R - P$, is a measure of the stakes of a Prisoner's Dilemma. It is one thing to face a decision between being in prison for one day and being in prison for two days; it is quite another to face a decision between being in prison for one day and being in prison for 50 years. The stakes of the latter decision are much greater than the stakes of the former decision. The first two derivatives entail that as the stakes of an interaction increase so too does the probability that an autonomous agent will cooperate. The third of the derivatives shows that as the temptation to defect in a Prisoners Dilemma increases, so too does the probability that an autonomous agent will defect.

The picture of an autonomous agent that emerges from a game theoretic analysis of autonomy in a Prisoner's Dilemma is thus a rather nuanced one. The fact that an autonomous agent's decision to cooperate is temptation dependent shows that autonomy does not entail the kind of supererogatory capacity always to avoid temptation that is entailed by Kant's Categorical Imperative. In this way, an autonomous agent is very much like ordinary human agents most of whom are not immune to the pull of temptation. Autonomous agents are like most ordinary humans in another way as well – they are sensitive to the stakes of their actions. Trivial actions that do not entail a great disadvantage to another do not weigh heavily on most people's conscience whereas actions that could greatly harm others do. The fact that autonomous agents are sensitive to the stakes of their actions provides the resources for a response to David Hume's famous assertion in his *Treatise of Human Nature*: 'It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger' (1978 [1739]: 415). Although Hume's claim may withstand scrutiny on certain conceptions of reason, the first two derivatives above provide a mathematical refutation of his claim if reason is understood in terms of the actions of an autonomous game-theoretic agent.

From the viewpoint of an autonomous agent, therefore, the correct assessment as to how one ought to act in a Prisoner's Dilemma is a complex affair, which seems appropriate given the tug of intuitions that the Prisoner's Dilemma often elicits. From the standpoint of a Nash reasoner, agents ought to defect. Nonetheless, there is a strong intuition that both players ought to cooperate, since such a decision leads to a Pareto and socially optimal outcome. We are not the first to note this battle of intuitions. This is what Gold and Sugden (2007) say about the matter:

The theory prescribes defect, but many people have the strong intuition that cooperate is the rational choice. Of course, it is open to the game theorist to argue that that intuition is mistaken, and to insist on the normative validity of the standard analysis. In doing so, the game theorist can point out that any individual player of the Prisoner's Dilemma does better by choosing defect than by choosing cooperate, irrespective of the behaviour of her opponent. In other words, each individual player can reason to the conclusion: 'The action that gives the best result for me is defect'. But, against that, it can be said with equal truth that the two players of the game both do better by their both choosing cooperate than by their both choosing defect. Thus, each player can also reason to the conclusion: 'The pair of actions that gives the best result for us is not (defect, defect).' It seems that

normative argument between these two positions leads to a stand-off. (Gold and Sugden 2007: 118)

From the viewpoint of an autonomous agent, both the intuitions that Gold and Sugden mention have merit, though there is room for an even more fine-grained analysis than they suggest. Not only can both cooperating and defecting be a rational response to a Prisoner's Dilemma, but the extent to which they are rational depends on the temptation and the stakes involved.

4. Psychic and material payoffs

In the last section we provided an analysis of the behaviour of an autonomous agent. Such behaviour stems from the set of psychic payoffs that are entailed by the universalizing operation. Although an autonomous moral agent chooses to behave in accordance with the universalizable equilibrium strategy that maximizes his expected psychic payoffs based on UPD, his actions nonetheless have material payoffs that are given by PD. This raises the question as to the difference between the psychic and material outcomes for an autonomous agent.

When an autonomous agent decides to use the universalizable equilibrium strategy of UPD in order to determine how to play PD, he enters an 'as-if' situation where the appropriate moral action is to play PD according to the universalizable strategy of UPD. Since the universalizable strategy entails the assumption that the other player plays the same strategy too, the psychic payoff that an autonomous moral agent receives from adopting the UPD framework equals:

$$\begin{aligned} & \left(\frac{R-P}{T-P}\right)\left(\frac{R-P}{T-P}\right)R + \left(\frac{R-P}{T-P}\right)\left(1 - \frac{R-P}{T-P}\right)R \\ & + \left(1 - \frac{R-P}{T-P}\right)\left(\frac{R-P}{T-P}\right)T + \left(1 - \frac{R-P}{T-P}\right)\left(1 - \frac{R-P}{T-P}\right)P = R \end{aligned}$$

It is interesting to note that the expected value of the psychic payoff of an autonomous moral agent based on the universalizable Nash equilibrium of UPD exactly equals the Pareto efficient material payoff an agent would receive in the cooperate-cooperate equilibrium of PD. This equivalence is a consequence of dealing with a mixed Nash equilibrium.

This mathematical result leads to one final philosophical takeaway from our account of autonomy. The relationship between acting morally and happiness is one of the oldest concerns in the history of ethics, stemming as far back as Plato's *Republic*. There have been many notable philosophers who have argued that acting morally is non-accidentally related to happiness and can lead to inner peace even in the presence of material loss. Indeed, in the *Republic* Plato goes as far as to argue that it is better to be perfectly just and having one's eyes gouged out on a rack than perfectly unjust and living a life of ease in a palace (Plato 2010 [375 BCE]). Although many philosophers have expressed deep scepticism about such views, the mathematical result just displayed lends them some support.

In a Prisoner's Dilemma, an autonomous moral agent receives as a reward for his autonomy a psychic payoff equal to the reward for mutual cooperation. To receive such a payoff an autonomous agent must have the capacity to universalize in the first sense we have discussed, i.e. must be able to choose his level of moral engagement, and must universalize in the second sense, i.e. be committed to adopting a symmetrical equilibrium strategy that, because of its symmetry, all agents can adopt. It may of course be beyond an ordinary human's ability to universalize in both of these senses, and so we make no empirical claims about the likely psychological effect of someone's trying to be moral. It may be that the happiness that accompanies moral virtue is an outcome that only noumenal agents can expect to achieve. Nonetheless, the mathematical fact just displayed demonstrates that an agent who succeeds in being fully autonomous will, as the old adage expresses, find virtue to be its own reward. Of course, as Plato long ago acknowledged, being just is no guarantee that one won't end up suffering on a rack. The psychic reward for acting autonomously can coincide with diminished material payoffs. As an examination of the material payoffs of an autonomous agent will show, our model confirms such a fact.

The expected material payoff for an autonomous agent playing PD depends on the strategy his counterpart plays. If an autonomous agent interacts with another autonomous agent, his expected material payoff, $EMPA(A)$, equals:

$$\begin{aligned} EMPA(A) &= \left(\frac{R-P}{T-P}\right)\left(\frac{R-P}{T-P}\right)R + \left(\frac{R-P}{T-P}\right)\left(1 - \frac{R-P}{T-P}\right)(S+T) \\ &\quad + \left(1 - \frac{R-P}{T-P}\right)\left(1 - \frac{R-P}{T-P}\right)P \\ &= R - \frac{(T-R)(R-P)(R-S)}{(T-P)^2} \end{aligned}$$

Since $T > R > P > S$, $EMPA(A) < R$. In material terms, an autonomous moral agent fares worse than an agent playing PD in which both agents always cooperate. Allowing moral agents to autonomously decide the extent to which they should cooperate results in agents occasionally defecting. Thus, the autonomous moral agents in this framework fare worse than the Kantian agents as modelled by Roemer (2010, 2015, 2019). It can be shown that $EMPA(A) > P$ if $(T-P)^2 > (T-R)(R-S)$. Consequently, for PD games where this condition holds, two autonomous moral agents playing each other would achieve material payoffs that are a Pareto improvement over the outcomes achieved by two non-autonomous Nashian agents playing the unique Nash Equilibrium.

If an autonomous moral agent plays PD with another agent who always cooperates, the expected material payoff of the autonomous agent, $EMPA(C)$, equals:

$$EMPA(C) = \left(\frac{R-P}{T-P}\right)R + \left(\frac{T-R}{T-P}\right)T.$$

The value of this material payoff lies between R and T , i.e. $R < EMPC(C) < T$. As can be expected, by occasionally defecting and occasionally cooperating against an agent who always cooperates, the autonomous agent receives a higher material payoff than the reward payoff, R , but a lower material payoff than the temptation payoff, T . The cooperating agent playing against the autonomous agent receives an expected material payoff, $EMPC(A)$, which equals:

$$EMPC(A) = \left(\frac{R-P}{T-P}\right)R + \left(\frac{T-R}{T-P}\right)S.$$

Because $EMPC(C) > EMPC(A)$, when autonomous agents play cooperators, the autonomous moral agents perform better materially than the cooperators.

Finally, when an autonomous moral agent plays PD with another agent who always defects, the expected material payoff of the autonomous agent, $EMPA(D)$, equals:

$$EMPA(D) = \left(\frac{R-P}{T-P}\right)S + \left(\frac{T-R}{T-P}\right)P.$$

The value of this material payoff lies between S and P , i.e. $S < EMPA(D) < P$. As can be expected, by occasionally defecting and occasionally cooperating against an agent who always defects, the autonomous agent manages to earn a higher material payoff than the sucker's payoff, S , but a lower material payoff than the punishment payoff, P . The defecting agent playing against the autonomous agent receives an expected material payoff, $EMPD(A)$, which equals:

$$EMPD(A) = \left(\frac{R-P}{T-P}\right)T + \left(\frac{T-R}{T-P}\right)P.$$

Since $EMPD(A) > EMPA(D)$, when autonomous agents play defectors, the autonomous moral agents perform worse materially than the defectors.

To sum up, then, even though autonomous agents earn a psychic payoff equal to the reward, R , regardless of whom they play, when they compete against other types, autonomous agents fare worse materially than the defectors they interact with but better than the cooperators they interact with. This outcome is not surprising in PD as defection is the dominant strategy while cooperation is the dominated strategy. When playing among themselves, autonomous agents perform worse in material terms than cooperators playing among themselves, but, if $(T-P)^2 > (T-R)(R-S)$, they perform better in material terms than defectors playing among themselves. One can see in these results a mathematical articulation of the relationship between the noumenal and the phenomenal realm. Noumenally, autonomous agents fare as well as one could reasonably hope for in a PD. Although they do not always cooperate, their pattern of cooperation and defection is calibrated so that they fare as well as two agents who always cooperate. It is as if the noumenal world provides a refuge of necessity for autonomous agents. And yet, the phenomenal fate of autonomous agents bears the mark of contingency that generally permeates the phenomenal realm. Whether an autonomous agent fares better materially than his immoral Nashian competitors depends in part on the nature

of his interactive counterparty and in part on the values inherent in the Prisoner's Dilemmas that he faces.

5. Conclusion

Game theory seems an unlikely place to find an account of deontology. The standard Nashian solution concept leads in well-studied cases to behaviour that is decidedly non-moral. In perhaps the most famous game – the Prisoner's Dilemma – defect, not cooperate, is the deliverance of Nashian will. Despite the seeming incongruence between it and deontology, game theorists have recently introduced deontological considerations into game theory by way of the Kantian counterfactual. The two extant types of model, one introduced by Roemer and one by Alger and Weibull, both contain impressive mathematical results and hence show the fecundity of studying deontology game theoretically. In this paper, we have proposed a third type of model that incorporates the Kantian counterfactual by way of a recursively defined action type that requires an agent to choose her level of universalized behaviour on the basis of endogenous features. Such a model, we contend, provides an analysis of autonomy.

As should be expected, our account of morality bears both similarities and dissimilarities to the accounts given by Kant, Roemer, and Alger and Weibull. We share with all three an approach to morality that stems from the formal structures involved in moral thought. In this way, we depart sharply from the currently widespread methodology within contemporary analytic ethics that takes its start from intuitions about empirically described cases: situations ranging from trolleys barrelling toward people (Thomson 1976), to children drowning in lakes (Singer 1972), to lecherous millionaires (Feinberg 1989) have all been raised to pump intuitions that are then marshalled for or against moral principles. Although we do not deny the relevance of such intuitions for moral theorizing – though we do note in passing that a methodology that relies on them must contend with well-known behavioural biases that threaten to seriously weaken their epistemic strength (Petrinovich and O'Neill 1996; Swain *et al.* 2008; Wright 2010; Cameron *et al.* 2013) – in this paper we have approached the moral ought by way of an a priori analysis of the fundamental moral concepts of universalizing and autonomy.

Though not a fully general mathematical treatment, the mathematical analysis in this paper nonetheless delivers a novel and telling rule of morality. In situations described by the Prisoner's Dilemma, an autonomous agent would engage in temptation and stakes dependent cooperative behaviour. In its stakes dependency such a rule contradicts the Humean principle that all fundamental desires, whether they concern one's little finger or the blowing up of the world, are equally rational. And in both its stakes and temptation dependencies, such a rule avoids the exacting demands of Kant's categorical imperative.

In addition to generating a rule of morality, the analysis in this paper has a philosophical implication that is deeper than any single rule of morality and that suggests novel avenues of inquiry. By analysing a fundamental moral notion in terms of a recursively defined action type, we have shown that a fundamental

moral decision, namely the decision to universalize one's actions, entails a set of psychic payoffs that, we contend, characterize an agent in the noumenal realm. In the *Critique of Pure Reason*, Kant (1996 [1781]) argued that the noumenal realm, consisting of objects as they are in themselves, is unknowable by human minds that must of necessity employ a set of a priori concepts. Contra Kant, we have suggested that a relation between the phenomenal and the noumenal, at least with respect to moral agents, can be specified recursively and so have suggested that there may be a mathematically precise, and hence knowable by human minds, articulation of the relationship between the two domains.

References

- Alger I. and J.W. Weibull** 2013. Homo moralis – preference evolution under incomplete information and assortative matching. *Econometrica* **81**, 2269–2302.
- Alger I. and J.W. Weibull** 2016. Evolution and Kantian morality. *Games and Economic Behavior* **98**, 56–67.
- Bacharach M.** 1997. *“We” Equilibria: A Variable Frame Theory of Cooperation*. Oxford: Institute of Economics and Statistics, University of Oxford.
- Binmore K.** 1994. *Game Theory and the Social Contract, Volume 1: Playing Fair*. Cambridge, MA: MIT Press.
- Bolton G. and A. Ockenfels** 2000. ERC — a theory of equity, reciprocity and competition. *American Economic Review* **90**, 166–193.
- Cameron C., K. Payne and J.M. Doris** 2013. Morality in high definition: emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology* **49**, 719–725.
- Chant S.R. and Z. Ernst** 2008. Epistemic conditions for collective action. *Mind* **117**, 549–573.
- Falk A. and U. Fischbacher** 2006. A theory of reciprocity. *Games and Economic Behavior* **54**, 293–315.
- Feinberg J.** 1989. *The Moral Limits of the Criminal Law: Volume 3: Harm to Self*. New York, NY: Oxford University Press.
- Gilbert M.** 1992. *On Social Facts*. London: Routledge.
- Gold N. and R. Sugden** 2007. Collective intentions and team agency. *Journal of Philosophy* **104**, 109–137.
- Grafton R.Q., T. Kompas and N.V. Long** 2017. A brave new world? Kantian–Nashian interaction and the dynamics of global climate change mitigation. *European Economic Review* **99**, 31–42.
- Hume D.** 1978 [1739]. *A Treatise of Human Nature*, ed. L.A. Bigge and P.H. Nidditch. Oxford: Oxford University Press.
- Kant I.** 1996 [1781]. *Critique of Pure Reason*, ed. J.W. Ellington, P. Kitcher and W.S. Pluhar. Indianapolis, IN: Hackett Publishing.
- McCann H.J. and M.E. Bratman** 1991. Intention, plans, and practical reason. *Noûs* **25**, 230.
- Newton J.** 2017. Shared intentions: the evolution of collaboration. *Games and Economic Behavior* **104**, 517–534.
- Petrinovich L. and P. O’Neill** 1996. Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology* **17**, 145–171.
- Plato** 2010 [375 BCE]. *The Republic of Plato*, Reissue Edition, ed. J. Adam. New York, NY: Cambridge University Press.
- Rabin M.** 1993. Incorporating fairness into game theory and economics. *American Economic Review* **83**, 1281–1302.
- Roemer J.E.** 2010. Kantian equilibrium. *Scandinavian Journal of Economics* **112**, 1–24.
- Roemer J.E.** 2015. Kantian optimization: a microfoundation for cooperation. *Journal of Public Economics* **127**, 45–57.
- Roemer J.E.** 2019. *How We Cooperate: A Theory of Kantian Optimization*. New Haven, CT: Yale University Press.
- Roth A.S.** 2004. Shared agency and contralateral commitments. *Philosophical Review* **113**, 359–410.

- Searle J.R.** 1990. Intentionality and its place in nature. In *Philosophy, Mind, and Cognitive Inquiry: Resources for Understanding Mental Processes*, ed. D.J. Cole, J.H. Fetzer and T.L. Rankin, 267–280. New York, NY: Springer Publishing.
- Sen A.** 1974. Choice, ordering and morality. In *Practical Reason*, ed. S. Körner, 54–67. Oxford: Blackwell.
- Sen A.** 1977. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6, 317–344.
- Singer P.** 1972. Famine, affluence, and morality. *Philosophy and Public Affairs* 1, 229–243.
- Smerilli A.** 2014. Theories of team reasoning. *Philosophical Readings* 6, 24–34.
- Swain S., J. Alexander and J.M. Weinberg** 2008. The instability of philosophical intuitions: running hot and cold on truetemp. *Philosophy and Phenomenological Research* 76, 138–155.
- Thomson J.J.** 1976. Killing, letting die, and the trolley problem. *Monist* 59, 204–217.
- Tuomela R. and K. Miller** 1988. We-intentions. *Philosophical Studies* 53, 367–389.
- Wallace R.J.** 2013. The deontic structure of morality. In *Thinking About Reasons: Themes from the Philosophy of Jonathan Dancy*, ed. D. Bakhurst, M.O. Little and B. Hooker, 137–168. Oxford: Oxford University Press.
- Wright J.C.** 2010. On intuitional stability: the clear, the strong, and the paradigmatic. *Cognition* 115, 491–503.

Paul Studtmann is Full Professor of Philosophy at Davidson College. He is the author of *The Foundations of Aristotle's Categorical Scheme* (Marquette University Press, 2008) and *Empiricism and the Problem of Metaphysics* (Lexington Books, 2010). His current research interests include the game-theoretic foundations of ethical reasoning and the set-theoretic foundations of theology. URL: <https://www.davidson.edu/people/paul-studtmann>.

Shyam Gouri Suresh is Associate Professor of Economics at Davidson College. He has published papers on a wide range of topics that include applications of agent-based modelling, macroeconomics methodology, opinion dynamics, empirical open economy macroeconomics, and the economics of migration. His current research interests include poverty traps in developing countries, the game theoretic foundations of ethical reasoning, and the co-evolution of economic policy, economic thought, and economic outcomes related to inequality and efficiency in the United States. More information is available on his website: <http://shyams.org/>. URL: <https://www.davidson.edu/people/shyam-gouri-suresh>.

Cite this article: Studtmann P and Gouri Suresh S (2021). Universalizing and the we: endogenous game theoretic deontology. *Economics & Philosophy* 37, 244–259. <https://doi.org/10.1017/S0266267120000279>