The use of evolutionary game theory to explain the evolution of human norms and the behavior of humans who act according to those norms is widespread. Both the aims and motivation for its use are clearly articulated by Harms and Skyrms (2008) in the following passage, which provides the focus for much of our ensuing discussion.

A good theory of evolution of norms might start by explaining the evolution of altruism in Prisoner’s Dilemma, of Stag Hunting, and of the equal split in the symmetric bargaining game. These are not well-explained by classical game theory based on rational choice. From a technical point of view, they present different theoretical challenges. In the bargaining game, there are an infinite number of equilibria with no principled (rational choice) way to select the cooperative one. In Stag Hunt there are only two, but the non-cooperative one is selected by risk-dominance. In Prisoner’s Dilemma the state of mutual cooperation is not a Nash equilibrium at all, and cooperation flies in the face of the rational-choice principle that one does not choose less rather than more. In contrast to rational choice theory, the most common tool of evolutionary game theory is the replicator dynamics, in which the propagation rate of each strategy is determined by its current payoffs. These dynamics have a rationale in both biological and cultural evolutionary modeling, and sometimes tell us things that rational choice theory does not.

We firmly agree with the first sentence in this quotation: a good theory about behavior under norms ought to explain altruism in the Prisoner’s Dilemma (PD), playing Stag in Stag Hunt (SH), and offering equal splits in the symmetric Nash bargaining game (NB). We also agree with Harms and Skyrms about the difference in technical challenges each of these games pose. Finding a single mechanism, even one as broadly understood as evolution, that could solve these challenges *en masse* is no doubt a tall order. Nonetheless, in this paper we present a single, simple, modification to SH, NB and a general n-player PD that does just that.

From a mathematical point of view the modification is less substantive than the modification involved in evolutionary game theory. In the first instance, we do not appeal to repeated games. All the models we examine in this paper are one-shot perfect information games. Moreover, in solving the models we appeal to the standard Nash equilibrium. Hence, we do not require the more substantive equilibrium concept of evolutionary stability. Finally, we do not incorporate any equations or theoretical apparatus from other sciences. This reduction in mathematical complexity does require an increase in the conceptual content of the models. This shouldn’t be surprising. Nothing comes for free. But the increase in conceptual content in this instance is directly related to the agents that are the target of the explanation. Unlike the models of classical game theory, our models directly incorporate the concept of autonomy. In contrast to the evolutionary explanation of human moral behavior, we call our explanation a *normative explanation*. We show that modeling agents as having a distinctive type of autonomy, what we call *deontological autonomy*, leads to the precise outcomes that evolutionary game theory has been trying to explain.

Our argument for this claim is a continuation of the argument begun in Studtmann and Gouri-Suresh (2021), who introduce a model of deontological autonomy and apply it to PD. Their model, what they call *Universalized Prisoner’s Dilemma* (UPD) is an instance of one of three types of model that have recently been introduced to capture what are called *Kantian* preferences, i.e., preferences for outcomes in which everyone acts the same. The other two occur in Roemer (2010, 2015, 2019), and Alger and Weibull (2013, 2016). The Studtmann/Gouri-Suresh models differ from the other two in that they endow agents with a choice to universalize their actions. In their models, situations that typically involve a single strategic choice are modeled as if there are two choices an agent must make: the standard strategic choice, and a choice as to whether one will make such a choice while motivated with selfish motivations or with the motivations inherent in a deontological rule of morality. Studtmann and Gouri-Suresh show that deontologically autonomous agents are altruistic. They are not perfectly altruistic – some amount of defection is inevitable. There is nonetheless always a non-zero probability that such agents cooperate in PD, and the probability increases as the stakes of the dilemma increase. Hence, Studtmann and Gouri-Suresh have shown that deontological autonomy solve the challenges posed by PD.

In this paper, we extend the Studtmann/Gouri-Suresh model to the Public Good Game (PGG), which is an n-player PD, the Stag Hunt (SH), and the symmetric Nash bargaining game (NB). Although the theoretical challenge posed by PGG does not differ from the theoretical challenge posed by PD – both are social dilemmas in which the Nash equilibrium is non-cooperation -- we include PGG because humans cooperate not just in dyadic interactions but in contributions to a public good being produced by very large groups. Indeed, if humans did not have the ability to cooperate in large groups for a common goal, it is hard to see that there would be a problem of moral norms at all. The move from a 2-player game to an n-player game is mathematically non-trivial. The inclusion of PGG in this paper can thus be seen as a direct continuation of the Studtmann/Gouri-Suresh treatment of PD. We demonstrate the existence of a unique symmetric equilibrium in the Universalized Public Good Game (UPGG) in which agents contribute to the public good in a way that is sensitive to the value of the public good – the greater the value of the public good the greater the probability of cooperation – and is largely independent of the number of agents playing the game. We move beyond the Studtmann/Gouri-Suresh argument, however, by showing that deontological autonomy also solves the challenges posed by SH and NB. We demonstrate that deontologically autonomous agents play Stag in the Universalized Stag Hunt (USH) and offer a 50-50 split in Universalized Nash Bargaining Game (UNB).

In addition to showing that deontological autonomy yields this kind of behavior, we discuss the general type to which the models of deontological autonomy belong. Gouri-Suresh and Studtmann (2022) discuss the general type extensively and provide a literature review of alternative attempts to incorporate morality into game theory. The discussion in this paper will be much more limited (and will omit a literature review) but will nonetheless show that the models of deontological autonomy that we employ are an instance of a very natural and very general type of model that we call an *Endogenized Morality Model (EMM).* EMM’s allow one to study within game theory a multitude of different types of autonomy. Deontological autonomy is one of several other natural types, for instance utilitarian autonomy and empathy-based autonomy. Each type of autonomy is distinctive both in its underlying moral principles and in the behavior of its agents. The choice of deontological autonomy as the proper explanation of human behavior is thus a substantive choice, one that is not forced by the adoption of the general type of model.

The fact that one can provide a mathematically precise normative explanation of the behavior of human agents has several theoretical implications. Among other things, it undermines the main motivation for the appeal to evolutionary game theory as a way of explaining human norms, which as Skyrms points out above is directly tied to the failure of classical game theory to explain cooperation in PD, Stag in SH, and equal splitting in NB. There is in addition a positive implication concerning the nature of human agency. The models in this paper assume that moral agents do not face just strategic choices but also face the choice as to how they view the values involved in a strategic choice. It is that choice, the choice to view the world through some set of values, we contend, that makes an agent autonomous. It is that choice that constitutes moral freedom. We show in this paper that one can model that choice as a strategic choice within game theory. One might call the resulting freedom *strategic* moral freedom. We show beyond any mathematical doubt that agents with deontological strategic moral freedom behave in precisely those ways that evolutionary game theorists have been trying to explain.

Several philosophers have argued that evolutionary and game-theoretic accounts of morality leave out crucial dimensions. Some of the most compelling cases come from those who emphasize the role of autonomy in human reasoning and action. (Fitzpatrick 2018, 2020 Buchanan and Powell 2015, Korsgaard 2006, Nagel 1979) We agree with these philosophers that evolutionary accounts of morality as well as game-theoretic accounts that do not incorporate autonomy are defective, not just conceptually, but also behaviorally. As Harms and Skyrms point out, classical game theoretic models make terrible predictions. And despite many impressive successes, evolutionary game theorists have yet to offer a complete explanation of the human tendency to play Stag, offer 50-50 splits, and cooperate in Public Good games. And of course, evolutionary explanations of these behaviors do not consider the normative features of the agents who are being modeled.

We show in this paper that by introducing deontological autonomy directly into game-theoretic models, the behavioral problem is solved. Deontologically autonomous agents play Stag, offer 50-50 splits, and cooperate in social dilemmas. Moreover, introducing deontological autonomy directly into game theoretic models also solves, at least to a significant degree, the conceptual deficiencies in game theoretic analyses of human morality. For, autonomy is arguably the fundamental concept in moral philosophy. Many philosophers, too many to list, both historical and contemporary, have argued that humans are moral agents who are moved by moral concerns. By incorporating autonomy directly into game theoretic models, we show that normative explanations of human behavior can be given a precise mathematical articulation and that by way of the standard Nash equilibrium they entail playing Stag in SH, offering 50-50 splits in NB, and cooperating in PGG.

The rest of our paper is structured as follows. In section I, we discuss EMM’s first generally and then by way of the construction of Universalized Stag Hunt (USH). In the remaining sections, we discuss one game per section. At the beginning of each section, we review briefly evidence that humans engage in behavior that calls for explanations of the sort that game theorists have been trying to provide. We also review the progress that evolutionary game theory has made in providing those explanations. These brief reviews are not meant to be exhaustive but are meant to highlight both the successes of evolutionary game theory and a difference between normative explanations and evolutionary explanations: The explanations that evolutionary game theorists offer are conditional, often highly so. This is of course recognized by evolutionary game theorists. Consider, for instance, Bendor and Swistak (2001) who argue for an evolutionary account of the existence of social norms:

Hence our result is robust under small perturbations of the parameters of the model. It is not, however, universal: it does not hold for all values of the parameters. Norms do not always replicate more easily. As with any formal result, we can only expect the deductive conclusions to be observed in reality if this reality conforms to the assumptions of our model.   
*\*

For a strategy to become evolutionarily stable, substantive empirical conditions must be in place. This does not by itself count against the explanations, since the conditions may be quite natural. But in some cases, natural conditions can be found that lead to alternative strategies in equilibrium. In such cases, it is questionable to what extent the explanations that are being given are in fact *explanations*. (D’ Arms, et al., 1998) The explanations we provide, on the other hand, are conditional upon only a single assumption, one that is mathematically simple to incorporate into game theory and one that in the case of humans possesses a great deal of phenomenological support, namely that the agents being modeled have deontological autonomy. We show that making that single assumption allows one to employ the classical Nash equilibrium to derive the intended strategies. In section II, we show that USH has Stag as the only strategy in equilibrium. In section III, we show that UNB has a 50-50 split as the only equilibrium. And, in section IV, we discuss the unique symmetric equilibrium in UPGG.

*Section I – Endogenized Morality Models*

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 first way depends on intuitions about the way the world ought to be. With this method, 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 a game that is transformed according to a deontological principle a *Deontology*. In a *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. There are other possible transformations that reflect other moral principles. For instance, one can transform PD according to a utilitarian principle. In such a transformed game, 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. In this paper, however, we confine ourselves to deontological morality.

In addition to the intuitive derivation of a morality, a morality should be derivable from a parameterization that captures the preference structure of an agent who has preferences in line with the moral theory. So, for instance, 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 , SH with the Alger-Weibull utility function reduces to a *Deontology*.

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 . We assume that the identity transformation yields a morality of materialistic selfishness. The transformed strategy space, , available to each moral agent is the Cartesian product of the morality space and the action strategy space, .

This approach to constructing models of autonomy is general and can be applied to any standard normal form non-cooperative game. Because of their generality, it is possible to construct EMM’s that represent a variety of autonomous agents. So, for instance, Marc Levine introduces two parameters, an altruism parameter, a, to model what can be called the *utilitarian* preferences agents have, and a a parameter l that represents an agent’s attitude about the other other agent’s altruistic preferences. When a=1 and l=0, Levine’s parameterization reduces to a utilitarian morality. An agent that has the capacity to choose a utilitarianism over some other morality, for instance a morality of materialistic selfishness, would have utilitarian autonomy. For any utility function that contains a parameter that can be consistently extremized, one can generate a corresponding morality. Gouri-Suresh and Studtmann (2022) discuss several such moralities. In this paper, we restrict our attention to autonomy in which an agent chooses between a morality of materialistic selfishness and a deontological morality that results from setting the Alger-Weibull parameter to 1.

To see the construction of Universalized Stag Hunt (USH), consider first the following matrix representation of the Stag Hunt, where R.

Table 1: Stag Hunt (SH)

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

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?’. A deontological version of SH results from transforming SH by a rule that in effect forces agents to live with the results of everyone acting like they do. According to the transformed game, an agent receives as a payoff what she would have received in the original SH had both players acted in the way that she is acting. If a player plays Stag, she receives what she would have received in SH had both players played Stag, and if she plays Hare, she receives what she would have received in the original SH had both players played Hare.

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 SH with the Alger-Weibull utility function in the context of pure strategies.

Table 3: Parameterized Stag Hunt

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

When , *Parameterized Stag Hunt* reduces to SH, 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.

With *Deontology* in place, it is straightforward to construct a matrix for Universalized Stag Hunt, which is as follows. (Notice that SH occurs in the bottom right 2x2 quadrant of the matrix and *Deontology* occurs in the upper left 2x2 quadrant. The off diagonals are split, so to speak, between the two.)

**Table 6:** *Universalized Stag Hunt* (USH)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | R, R | R, P | R, R | R, T |
|  | P, R | P, P | P, S | P, P |
|  | R, R | S, P | R, R | S, T |
|  | T, R | P, P | T, S | P, P |

*Section II – Universalized Stag Hunt*

Humans have a remarkable ability to coordinate with each other to accomplish mutually advantageous goals. The standard type of game to model such situations is a coordination game. Two early sets of experimental results were interpreted as suggesting that coordination failure in laboratory experiments is common (???). Bortolotti, et. al. (2016) show, however, that if allowed to play repeatedly laboratory subjects coordinate about 80% of the time. Devetag and Ortman (2020) have identified the major determinants of such failures and argue that actual coordination failures are likely to be the exception rather than the rule in both laboratory experiments and in real world settings. One of the most common forms of coordination consists in turn taking. Grueneisen and Tomasello (2017) have shown that human children as young as three years old engage in turn-taking as a solution to a coordination problem. Melis et al. (2016) have shown that by the age of five human children have a stable ability to take turns in sharing resources, which is not an ability that chimpanzess have. Nonetheless, turn taking is not unique to humans. Martin, et. al., (2017) have shown that chimpanzees are able to take turns to solve experimentally induced number sequencing tasks. The ability in humans, however, is far more highly developed and integral to human social arrangements. Skyrms (2001) argues that despite all the attention devoted by evolutionary game theorists to explaining cooperation in the Prisoner’s Dilemma the evolution of coordination in Stag Hunt deserves at least as much attention.

Evolutionary game theory has had success in this area. A general result of evolutionary game theory is that assortative pairing – when cooperators (coordinators) interact with cooperators (coordinators) and defectors interact with defectors – cooperative strategies become evolutionary stable. (Tarnita, ???) Zollman (2005) shows that a local interaction model combined with signaling leads to agents playing Stag. Skyms and Pemantle (2000) show that Stag can flourish if agents are capable of learning. Ely (2002), Oechssler (1999), and Diekemann (1999) show that if Stag hunters can move to islands where they are isolated, Stag can stabilize. The type of rationality that the agents have also makes a difference in the evolution of Stag. Antoci, et. al. (2000), show that correlated equilibrium allows Stag to stabilize if players are boundedly rational. Nokelby and Stirling (2006) show that Stag can emerge via replicator dynamics if the players are satisficers.

These results illustrate a typical feature of evolutionary game-theoretic explanations of the emergence of human norms, namely their multiple conditionality. Several different conditions have been specified that allow Stag to become a stable strategy. One thus naturally wonders which condition explains its evolution in the human case. Did humans evolve to play Stag because they are satisficers, or because human Stag hunters became isolated, or because humans are signalers, or because humans are capable of learning? Moreover, one wonders whether the same condition that explains the evolution of playing Stag in humans also explains the evolution of 50-50 splits in NB and cooperation in PD. As far as evolutionary game-theoretic explanations go, it is certainly possible that humans are hard-wired to play Stag as a result of a process involving one condition and hard-wired to offer 50-50 splits as the result of a different process involving different condition. It seems plausible, however, that there is some connection between the behaviors. It seems plausible that the human tendencies to play Stag, offer 50-50 splits, and cooperate in social dilemmas are manifestations of a single underlying evolved capacity. Thinking of each of the behaviors as independently hard-wired ignores the role that human reason, more specifically human *moral* reasoning, might play in the behaviors. If anything seems hard-wired into humans, it is the capacity to reason, especially morally, about situations and then to be moved by such reasons.

These features of evolutionary explanations stand in contrast to the normative explanation. On the assumption that agents have deontological autonomy, it follows directly from the ordinary Nash equilibrium that playing Stag is the only strategy in equilibrium, which can easily be shown. In SH, one assumes . With such an ordering, US in the above matrix strictly dominates UH and ~UH. Hence, the above matrix reduces to the following matrix.

**Table 6:** *USH Reduced*

|  |  |  |
| --- | --- | --- |
|  | US | ~US |
| US | R, R | R, R |
| ~US | R, R | R, R |

In this matrix, Stag is the only possible strategy. Hence, deontologically autonomous agents play Stag whether they universalize their actions or not. In other words, the mere possibility of universalizing suffices to generate Stag as an equilibrium. One would thus expect deontologically autonomous agents to be very good at coordinating their behavior.

*Section III – Universalized Bargaining*

Humans have egalitarian preferences that are arguably related to the emergence of egalitarian food sharing in hunter gatherer societies. Anthropological studies have shown food sharing practices to be a universal feature of hunter gatherer societies, though the specific details of the sharing practices differ with some societies being more explicitly egalitarian and others allowing differences in sharing due both to genetic and social relations and to notions of desert. (Barnard 2002, Dowling 1968, Leacock and Lee 1982, Lee and Devore 1968, Service 1966) Chimpanzees do not share food with non-kin, though both bonobos (Hare, Kwetuenda 2010) and capuchin monkeys (de Waal 2000) do so to a limited degree.

The Nash bargaining game (NB) can be used to study the emergence of egalitarian preferences. In the Nash bargaining game, strategies are represented by a pair (*x*, *y*). where *x* and *y* are selected from the interval [*d*, *z*], where *z* is the total good. If *x* + *y* is equal to or less than *z*, the first player receives *x* and the second *y*. Otherwise both get d, which for ease of presentation we assume equals zero. For any *x* and *y* such that *x* + *y* = *z*, (*x*, *y*) is a Nash equilibrium. This means that highly inegalitarian outcomes are Nash equilibria in NB, where the egalitarian outcome would be a 50-50 split.

As in the case of Stag Hunt, evolutionary game theory has made significant progress in studying the evolution of 50-50 splits in NB. Skyrms (1996) offers a model using the replicator dynamics of Taylor and Jonker (1978) that contains natural conditions under which 50-50 splits become stable. D' Arms et al. (1998) criticize Skyrms’s account and describe conditions that in Skyrms’s model leads to the stability of one agent offering 1/3 and the other agent offering 2/3. Alexander (2000), however, contains a spatial model of divide-the-dollar that has more robust convergence properties than the models of Skyrms and that avoids the methodological criticism in D' Arms et al.

There have been several attempts to explain 50-50 splits without appeal to evolutionary game theory. Of particular interest is Suleiman (2018) who modifies the standard utility function by introducing an epistemic factor that represents an agent’s aspirational level. Suleiman’s approach replaces the Nash equilibrium and its refinements with an alternative concept of ‘harmony’ in which the intersection of strategies leaves all agents equally satisfied. Suleiman shows that his model yields 50-50 splits in NB. Unlike Suleiman, the model of deontological agency in this paper relies on the Nash equilibrium. If one abstracts from the technical details, however, there is arguably a conceptual connection between Suleiman’s notion of harmony and universalizing – an agent who universalizes her actions chooses a motivational scheme in which everyone acts according to the same strategy. In such a situation, payoffs are harmonized.

As in the case of Stag Hunt, the normative explanation of 50-50 splits is conditional only upon the assumption that those playing NB are deontologically autonomous. With such an assumption, the only equilibrium is a 50-50 split. An informal argument for this conclusion can be easily given. Consider two agents who can universalize their decision so that they receive as a payoff what they would receive were the other agent to bid as they do. And suppose that . In any such game, any bid less than 50 is strictly dominated by the decision to universalize and bid 50. Hence, the universalized bargaining game reduces to a game in which the bids must be greater than or equal to 50. But in any such game, any bid greater than 50 is strictly dominated by the decision to universalize and bid 50. Hence, autonomous agents would bid 50 in the Nash bargaining game.

A formal presentation of the argument is as follows. Suppose that there are three possible bids: low, L, equal, E, and High, H. And assume that a bid of H+H is greater than the total good, bids of L+L, L+E, and L+H are less than or equal to the total good, and a bid of E+E equals the total good. (The assumption that L+H is less than or equal to the total good is made here for ease of presentation – it could be relaxed.) Then, the following matrix gives the payoffs.

**Table 6:** *Universalized Bargaining Game (UBG)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | L, L | L, E | L, 0 | L, L | L, E | L, H |
|  | E, L | E, E | E, 0 | E, L | E, E | E, 0 |
|  | 0, L | 0, E | 0, 0 | 0, L | 0, 0 | 0, 0 |
|  | L, L | L, E | L, 0 | L, L | L, E | L, H |
|  | E, L | E, E | 0, 0 | E, L | E, E | 0, 0 |
|  | H, L | 0, E | 0, 0 | H, L | 0, 0 | 0, 0 |

In this matrix, UE strictly dominates UL, UH and ~UL. Hence, it reduces to the following matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 7:** *UBG Reduced* |  |  |  |
|  | E, E | E, E | E, 0 |
|  | E, E | E, E | 0, 0 |
|  | 0, E | 0, 0 | 0, 0 |

In this matrix, UE strictly dominates ~UH. Hence, it reduces to the following matrix.

**Table 8:** UBG Reduced Further

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | E, E | E, E |
|  | E, E | E, E |

As in the case of Stag, the mere possibility of universalizing suffices for agents playing NB to offer 50-50 splits. Actual universalizing is unnecessary. One would thus expect deontologically autonomous agents to have very strong egalitarian tendencies.

*Section IV -- Universalized Public Good Game*

Humans, unlike chimpanzees, cooperate in large groups of people to provide the public good. Laboratory experiments have shown that the willingness to contribute to the public good is robust though partial. Empirical evidence on group size is mixed. Some studies such as Isaac et al. (1994) find that contributions increase with group size while other studies such

as Nosenzo et al. (2013) find that under some circumstances, contributions decrease with group size. Real world data suggests that humans are capable of sustained partial contribution to the public good in very large groups without a centralized authority. (Matthew, Boyd 2011)

One aspect of the data is particularly peculiar from a classical perspective. Unconditional contributions to the public good depend on the multiplier (the marginal per capita rate of return) – as the multiplier increases, so too does the level of contribution. (Fischbacher 2012) This runs counter to classic economic models according to which the rational strategy is independent of the multiplier.

Several interesting results have been obtained concerning the evolution of contribution to the public good. Bramoullé and Kranton (2007) show that social networks lead to equilibria in which some agents contribute to the public good and others do not. Brockhust et al. (2008) show that resource prevalence can influence contribution to the public good. Kurokawa and Ihara (2009) show that direct reciprocity obeying a one-third law can lead to stable outcomes in which agents contribute to the public good. Salahshour (2021) shows that consistent personalities across rounds can lead to contribution becoming stable. Once again, we note both the interest and multi conditionality of these results. For deontologically autonomous agents, on the other hand, contributing to the public good is a Nash equilibrium.

In the public good game that we discuss there are two levels of investment: zero, corresponding to defectors withholding their money, or a fixed amount, c, denoting the cooperator’s contribution. Without loss of generality, we assume that . The contribution of each player is multiplied by a multiplication factor r and added to the common pool, which is divided by the number of players N and distributed equally. For , the game is a PGG. The following matrix shows the payoffs in a universalized PGG for a single individual’s contribution to the public good when different numbers of other players are also contributing. We consider the situations in which 0, , and other players contribute.

**Table 4:** *UPGG*

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

Because , UC strictly dominates UD. Moreover, for column , because ; for column , because ; and for column because . Hence, ~UD strictly dominates ~UC. This matrix, therefore, reduces to the following matrix.

**Table 5:** *UPGG Reduced*

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

In this game, there are many asymmetric equilibria and a unique symmetric mixed-strategy equilibrium. We shall focus here on the symmetric equilibrium. The equilibrium is a mixture of UC and ~UD. In a symmetric equilibrium, each player cooperates with the same probability. Hence, we can assume that for all players and . The expected value of playing UC is r because that is the payoff regardless of what other players do. To find the expected value of playing ~UD, we need to recognize that the player with this strategy does not contribute to the pool and is able to keep his money which equals 1. In addition, this defector also enjoys the public good into which each of the remaining (N-1) players have contributed with a probability p. The expected values of the two strategies are therefore given by the following equations:

Setting the two expected values equal and solving for yields the following equation for the probability that an autonomous agent contributes to the public good, PR(UC):

From this equation one can see that when , and when approaches , approaches . When , increases monotonically with increases in . The limit of the probability function as goes to infinity is given by the following:

Hence, for high enough it is possible to maintain high levels of contribution to the public good even among very large groups of agents. The following is a plot of the probability that an agent contributes to the public good relative to increases in r for N = 20.



Autonomous agents thus act in a public good game like humans – as the multiplier increases so too does the willingness to contribute to the public good. Moreover, the cooperation of an autonomous agent is independent of genetic relatedness. Hence, one would expect autonomous agents to be able to cooperate in large groups of genetically unrelated individuals in n-player social dilemmas.

Although deontologically autonomous agents contribute to the public good, autonomy provides only a partial solution to the free rider problem, since deontologically autonomous agents also free ride. Universal contribution to the public good, if it is to be had at all, would thus require some other enforcement mechanism. The willingness to punish free riders and the effectiveness of punishment as a means for enforcing contribution to the public good has been well documented. Both in laboratory experiments and in real world settings, humans punish other agents, often out of a sense of fairness, who either fail to contribute to the public good or fail to cooperate in a social dilemma. (Fehr, Gachter 2000, 2002; Rockenbach, Millinski 2006; Weissner 2005, Matthew and Boyd 2011) The willingness to punish raises a theoretical puzzle, since punishing is subject to a second-order free-rider problem (Shinada, Yamagishi 2007). Boyd et al. (2003) have demonstrated by way of simulation that punishing behavior can become evolutionarily stable with the appropriate levels of group selection and agent imitation.

Although we once again do not dispute these findings concerning the evolution of punishment, we contend that group selection and agent imitation are not needed to explain why deontologically autonomous agents would punish. Because autonomy solves the first-order free rider problem to some degree, it would presumably solve at least to that same degree the second-order free-rider problem. Indeed, one could think of punishment as just another public good, in other words as presenting a first-order public good problem. Then, deontologically autonomous agents would be just as willing to punish as they would be to contribute to the public good in any other way. Moreover, because autonomous agents are motivated at least in part by a commitment to universalizable, and hence morally inflected, rules of action, one would expect that punishment of free riders would be associated with the sense that free riders *deserve* to be punished.

Conclusion

In this paper, we have examined deontological autonomy in three different games: Stag Hunt, the Nash bargaining game, and the Public Good game. We have shown that deontologically autonomous agents play Stag in Stag Hunt, offer 50-50 splits in the Nash bargaining game, and contribute to the public good. In addition, we discussed the general type of model, EMM’s, to which the models in this paper belong. Universalized games are EMM’s in which the underlying morality is one based on the universalizability of one’s actions.

It is somewhat surprising, though instructive, that the same modification to each of the three games would yield expedient results in each. For, recall what Harms and Skyrms say about the difference in the theoretical challenges that each raise:

“From a technical point of view, they present different theoretical challenges. In the bargaining game, there are an infinite number of equilibria with no principled (rational choice) way to select the cooperative one. In Stag Hunt there are only two, but the non-cooperative one is selected by risk-dominance. In Prisoner’s Dilemma the state of mutual cooperation is not a Nash equilibrium at all, and cooperation flies in the face of the rational-choice principle that one does not choose less rather than more.”

Despite the significant differences in the theoretical challenges posed by the three different games, the addition of deontological autonomy solves completely the challenges that two of the games present and partially, though in a realistic way, the challenges that the third game presents. Such a fact, we contend, explains why deontological autonomy would evolve. One thing that has been overwhelmingly demonstrated by evolutionary game theory is that playing Stag, offering 50-50 splits, and contributing to the public good *can* become stable strategies under various conditions. Given the evolutionary benefits that would arise for individuals of a species all of whom have some mechanism that could in a coordinated fashion solve the challenges faced by all three games, the results of this paper along with the results from evolutionary game theory suggest that with enough time and energy the evolution of deontological autonomy is all but inevitable.