The Evolution of Autonomy

**Abstract**

In this paper, we present a game-theoretic argument that humans evolved to cooperate among large groups of genetically unrelated individuals and to have preferences for an egalitarian distribution of surplus because they evolved into autonomous agents. We take as our understanding of autonomy what both Rousseau and Kant meant, namely that autonomy is obedience to a law that one has prescribed to oneself. The models we use for our argument are one of three types of models that have recently been introduced into game theory as a way of understanding Kantian morality. The other two occur in Roemer (2010, 2015, 2019), and Alger and Weibull ((2013, 2016). The approach we adopt appears in Studtmann and Gouri-Suresh (2021).

Unlike any other known species, humans cooperate in large groups of genetically unrelated individuals even in single encounters in which reputation gains are minimal, and they have tendencies toward egalitarianism. Each of these characteristics present an evolutionary puzzle and several hypotheses have been put forward to explain them. It is the purpose of this paper to present a game-theoretic argument that humans cooperate as they do and have the tendency toward egalitarianism that they do in part because they evolved into autonomous agents. It is not our intention to advance a novel understanding of autonomy or to add to the existing philosophical debates about it. Instead, we shall appeal to an understanding of autonomy that occurs in Rousseau and Kant whose views often serve as a historical and philosophical starting point for current discussions. According to Rousseau (1762), freedom is obedience to a law that one has prescribed for oneself. According to Kant (1785), a will is positively free if it can act according to a law of its own. It is the purpose of this paper to present game theoretic models that incorporate the idea that an agent can prescribe a universalized law to herself and to show that such a capacity can explain the above-mentioned human characteristics.

The models we use for our argument are one of three types of models that have recently been introduced into game theory as a way of understanding Kantian morality. The other two occur in Roemer (2010, 2015, 2019), and Alger and Weibull (2013, 2016). The approach we adopt appears in Studtmann and Gouri-Suresh (2021). The Studtmann Gouri-Suresh approach differs from the other two approaches insofar as they endow agents with a choice to universalize their actions. By universalizing her decision, an agent alters the payoffs involved in a game-theoretic situation, which has the effect of giving agents the motivation to be obedient to a rule of action that they have chosen. The assumption that agents maximize the expected value of a *universalized game* allows the models to be solved by way of the ordinary Nash solution concept.

In what follows, we discuss Nash equilibria that result from adding universalizing as an action type to three different games: The Prisoner’s Dilemma, The Public Good Game, and the Nash bargaining game. In their 2021 paper, Studtmann and Gouri-Suresh discuss the Universalized Prisoner’s Dilemma (UPD) as a way of understanding normativity. In this paper, we consider UPD as well as the universalized versions of the other two games predictively. We have chosen to look at the three games we have mentioned because we take them to be particularly relevant to question of the evolution of human behavior. For, it is humans’ ability to cooperate in social dilemmas and their tendency toward egalitarian norms of distribution that creates the evolutionary puzzle: how could the logic of evolution allow such behavior? How can the tendency toward free-riding and selfishness be curbed, if not perfectly, at least to a level that would allow for human cooperation to take the form that it does?

Our aim in this paper is to argue that one possible explanation for the pattern of human cooperative behavior is that humans evolved into autonomous agents. Our argument stems from the Nash equilibria of the three models we examine. From the equilibria, one would expect autonomous agents to: (1) adopt cooperative turn-taking as a solution to low-stakes dyadic social dilemmas; (2) adopt symmetrical stakes and temptation dependent mixed strategies in dyadic high-stakes social dilemmas; (3) contribute to the public good with a probability that increases as the value of the public good increases and that is largely independent of the number of agents in the population; and (4) share surplus equally. Importantly, autonomous agents would engage in all these behaviors with genetically unrelated individuals in one-shot anonymous encounters.

The very description of the behavior of autonomous agents should suffice to establish at least a general resemblance to human behavior. But of course, it is one thing to argue that the patterns of human behavior could be explained by the assumption that humans evolved into autonomous agents and another thing to argue that such an explanation is the correct one. Although trying to make a case for the correctness of the hypothesis is well beyond the scope of this paper, in the conclusion we present some considerations that count in its favor.

*Section I – Universalized Prisoner’s Dilemma*

We begin with the following standard representation of a Prisoner’s Dilemma.

**Table 1**: *Prisoner’s Dilemma (PD)*

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | R, R | S, T |
|  | T, S | D, D |

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 cooperator 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 .

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 universalizes, he receives as a payoff what he would have received in the original game had everyone played the strategy he is playing. Table 2 presents the result of adding universalizing to PD.

**Table 2**: *Universalized Prisoner’s Dilemma (UPD)*

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

One can see from this payoff matrix that universalizing changes the payoffs in the un-universalized PD. Consider for instance the square for UC/UD. In that outcome agent 1 receives a payoff of R even though he cooperates against a defector. In the original game, an agent who cooperates against a defector receives S. Hence, the act of universalizing alters the payoffs in the original game. Moreover, the alteration leads rational agents to cooperate. For UC strictly dominates ~UD. Hence, a rational agent who universalizes her action cooperates. By adding universalizing as an action type to a game, one in effect creates a kind of idealization within an agent’s motivational scheme. An agent who universalizes her cooperative action is motivated by the situation in which everyone cooperates even if not everyone is cooperating. In this way an autonomous agent can be motivated to choose the moral law.

With this model, it becomes appropriate to consider two different expected values: the expected value of playing UPD and the expected value of playing PD when using strategies from UPD. One can call the former value *agentive* *value* and the latter value *material value*. The agentive value includes the set of transformed values that may not match the *de facto* material outcomes. The material value arises from agents playing the original un-universalized games. We assume that autonomous agents act according to agentive value, but that evolution acts according to material value. The models we propose, then, are like those in Rotemberg (2008) who allows agents a choice that can alter their payoffs. Rotemberg, however, assumes that agents maximize material value whereas we assume that agents maximize agentive value but must live with the material outcomes that result.

To see how autonomous agents would behave, it is helpful to note that ~UD strictly dominates ~UC. Because UC strictly dominates UD, the above matrix reduces to the following matrix:

**Table 3:** *UPD Reduced*

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

This game has three Nash equilibria. Two are asymmetric pure-strategy equilibria and one is a symmetric mixed strategy equilibrium.

The asymmetric equilibria are UC/~UD and ~UD/UC. Let us define cooperative turn-taking as repeated playing of a game in which the players alternate between cooperate and defect an equal number of times. In the Prisoner’s Dilemma, cooperative turn taking would not involve any Nash equilibria, since defect/defect is the only equilibrium. In the above game, on the other hand, cooperative turn taking would always involve a Nash equilibrium. Moreover, in the original Prisoner’s Dilemma, repeated turn taking would yield an expected payoff for both players equal to , whereas in the above game repeated turn-taking would have an expected agentive payoff equal to 2. Because autonomous agents try to maximize agentive value, and because is not only greater than P but is also greater than R, from the point of view of agentive value, cooperative turn-taking is a Pareto improvement over mutual cooperation. Moreover, from the point of view of agentive value, there is little risk in cooperating first with the possibility that a less than scrupulous turn-taker does not cooperate next. For, the agentive value of cooperating against a defector is the same as the agentive value of cooperating against a cooperator. Of course, there is material risk. But being willing to take some material risk in the face of possible defection is one of the characteristics of autonomous agents. Indeed, such a tendency becomes quite vivid in the symmetric equilibrium, to which we now turn.

The symmetrical mixed-strategy equilibrium has a greater expected material payoff than cooperative turn taking when the stakes of the social dilemma become large enough. Hence, whatever agentive value there might be from cooperative turn-taking, because evolution works on material value, one would expect autonomous agents to adopt the mixed strategy for high-stakes social dilemmas. The mixed strategy is a mixture of UC and ~UD. The probability that an agent plays UC, Pr(UC), is given by the following equation.

The partial derivatives of with respect to each of its variables provide the basic pattern of behavior of an autonomous agent.

Given that , 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 behavior. In a Prisoner’s Dilemma, the difference between the reward for cooperation and the punishment for defection, , is a measure of the stakes of the dilemma. The first two derivatives entail that as the stakes of the PD increase so too does the probability that an autonomous agent cooperates. The third of the derivatives shows that as the temptation to defect in a PD increases, so too does the probability that an autonomous agent defects.

The expected agentive value of two autonomous agents playing the symmetrical equilibrium is R. Hence, from the point of view of agentive value, the symmetrical equilibrium is just as good as mutual cooperation. The expected material, as opposed to agentive, payoff of two autonomous agents playing the symmetrical equilibrium is given by the following equation:

if . Consequently, for PD games where this condition holds, two autonomous agents playing each other would achieve material payoffs that are a Pareto improvement over the outcomes achieved by two non-autonomous Nashian agents.

These facts show that autonomy could evolve. Interactions in a population are either perfectly assorted, i.e., autonomous agents always interact with autonomous agents and non-autonomous agents always interact with non-autonomous agents, or they are not. If a population is perfectly assorted, the above results show that autonomous agents will outcompete non-autonomous agents. As Tarnita, et al., have shown (2009), in populations that are not perfectly assorted, a more cooperative strategy is evolutionarily stable against a less cooperative strategy if agents playing the more cooperative strategy are more likely to interact with each other than with non-cooperative agents. Hence, with the appropriate population structure autonomy could evolve.[[1]](#footnote-1)

*Section II -- Universalized Public Goods Game*

The PGG is similar to an n-person PD. Hence, one would expect the behavior of autonomous agents in a PD and a PGG to be similar. And indeed, the behaviors are similar.

In the public good game we discuss, there only 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 symmetrical 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 agentive 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 contribution to the public good relative to increases in r for N = 20.



The behavior of autonomous agents in the symmetric equilibrium of UPGG is analogous to their behavior in the symmetric equilibrium in the UPD. In both cases, as the stakes increase, so too does the probability that an autonomous agent cooperates. In UPD, the stakes are determined by the difference between R and P. In UPGG, the stakes are determined by multiplication factor, r. It is also worth pointing out that the cooperation of an autonomous agent so far described is independent of genetic relatedness. Hence, one would expect autonomous agents to be able to cooperate in large groups of genetically unrelated individuals in both dyadic social dilemmas and in order to provide the public good.

One last point is worth making concerning the public good. It has been well-documented that 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) This willingness to punish has raised a theoretical puzzle, since punishing is subject to a second-order free-rider problem. (Shinada, Yamagishi 2007) But autonomy solves the first-order free rider problem. It doesn’t solve it completely, since for any multiplication factor, r, there is some level of free riding. Nonetheless, autonomous agents do contribute to the public good and do so more readily as the value of the public good increases. If autonomy solves the first-order free rider problem to some specifiable degree, then it presumably would solve at least to that same degree the second-order free-rider problem with respect to punishment. Moreover, given that 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.

*Section III – Universalized Bargaining*

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.

An informal argument that the only Nash equilibrium in the universalized game is an equal split 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 this argument can be given 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 |

It is interesting to note that in the UBG just allowing for the possibility of universalizing results in all agents playing cooperatively regardless of whether they universalize. In the UPD and the UPGG, on the other hand, the possibility of defecting continues to exist in the mixed strategy equilibrium and is paired with the decision not to universalize. One would thus expect autonomous agents to have very strong egalitarian tendencies.

*Section IV – Conclusion*

Based on three different game-theoretic models we have so far argued that autonomy could evolve and that if it did evolve autonomous agents would act very much like human agents have acted – they would cooperate with genetically unrelated individuals in social dilemmas and would share surplus in an egalitarian way. Hence, one possible explanation for human cooperative and egalitarian tendencies is that humans evolved into autonomous agents. An argument that such an explanation is not just a possible explanation but is in fact correct is beyond the scope of this paper. Nonetheless, we conclude by briefly considering a few reasons that count in its favor.

Because autonomy consists in a set of capacities that an agent can have or lack, it is the sort of characteristic that can be selected for. Moreover, it is plausible to suppose that the ability to decide in a motivationally effective way to obey a social rule would have given agents a selective advantage during a process of self-domestication, a process for which there is considerable evidence. (Zanella, et al., 2019) In addition to being consistent with methodological individualism, such a hypothesis also coheres with the view that group selection was a significant evolutionary force in human history. For, it is plausible to suppose that groups of agents who are willing to cooperate with genetically unrelated individuals in order to produce the public good would outcompete groups of non-autonomous agents who are not willing to do so. Another important aspect of the explanation is that it conforms to a rather plausible picture of human agents, namely that they are motivated both by their own material welfare as well as by considerations of the public good. The concern for the public good results from a kind of idealization in an autonomous agent’s motivational scheme, one that is independent of the actual course of events. An autonomous agent’s motivational scheme is partly decoupled from the de facto material payoffs that she will realize when she acts.  But it is only partly decoupled, since in addition to being motivated by their conception of the moral law, autonomous agents are also motivated by their desire for the *de facto* *material* payoffs that result from their actions. This can be seen most vividly in the symmetrical equilibrium in the UPD. Agents playing according to such a strategy cooperate less as the temptation to defect increases but cooperate more as the stakes of the dilemma increase. Thus, just like humans, autonomous agents are pulled in two different directions, toward their selfish and their pro-socialaims. In addition to explaining the willingness to cooperate, autonomy can also explain the malleability of human patterns of cooperation. Although the models of autonomy we have presented are highly idealized, in the real-world autonomy would be expressed relative to a recognized social code that an agent accepts as providing the correct norm for some context. An agent can decide that she should treat her children kindlier, she can decide that she should become a vegetarian, she can decide that she should join the military, and so on. Each of these decisions can motivate her to act as she has decided she ought to act. Although the specific nature of the rule she decides to follow differs from context to context, the basic pattern of behavior is the same: In each case, she obeys a rule that she has prescribed to herself.

There is one final, philosophical, reason to think that humans evolved into autonomous agents. A great deal of attention has been given to the question of the evolution of morality and whether the fact (if it is a fact) that morality evolved somehow debunks or undermines moral claims. (Street 2006, Joyce 2006, Fitzpatrick 2014, Fraser 2014, Rini 2016) Considerably less attention has been given to the question of the evolution of autonomy. (Korsgaard 2006, Singer 2011) Yet many philosophers hold that autonomy is necessary for moral agency. If so, an investigation into the evolution of autonomy is required for a complete investigation into the evolution of morality. In this paper, we have argued that it is possible for autonomy to evolve. If it is not just possible that autonomy evolved, but autonomy in fact figures into the correct explanation of human evolutionary history, then this necessary condition holds as a result of human evolution. And that, we contend, is both remarkable and philosophically significant. For, if autonomy did evolve, humans are not moral agents accidentally. Rather, moral agency is in our genes.

**Works Cited**

Alger, Ingela, and Jörgen W. Weibull .“Homo Moralis--Preference Evolution Under Incomplete Information and Assortative Matching.” Econometrica, vol. 81, no. 6, 2013, pp. 2269–2302., doi:10.3982/ecta10637.

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

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." American Economic Review, 90 (4): 980-994.

Fitzpatrick, William. 2014a, “Debunking Evolutionary Debunking of Ethical Realism,” Philosophical Studies, doi:10.1007/s11098-014-0295-y.

Fraser, B.J., 2014, “Evolutionary Debunking Arguments and the Reliability of Moral Cognition,” Philosophical Studies, 168: 457–473.

Joyce, R., 2006, The Evolution of Morality, Cambridge, MA: MIT Press.

Korsggard, Christine. “Reflections on the Evolution of Morality”, in *Primates and Philosophers: How morality Evolved*, Frans de Waal, 2006, Princeton University Press.

Kant, I., Bennett, C., Saunders, J., & Stern, R. (2019). *Groundwork for the metaphysics of morals*. Oxford University Press.

Matthew, Sarah, and Robert Boyd. “Punishment sustains large scale cooperation in prestate warfare.” PNAS July 12, 2011 108 (28) 11375-11380; https://doi.org/10.1073/pnas.1105604108.

Rini, R., 2016, “Debunking Debunking: A Regress Challenge for Psychological Threats to Moral Judgment,” Philosophical Studies, 173(3): 675–697.

Rockenbach, B., Milinski, M. “The efficient interaction of indirect reciprocity and costly punishment.” *Nature* **444,** 718–723 (2006). https://doi.org/10.1038/nature05229

Roemer, John E. “Kantian Equilibrium.” Scandinavian Journal of Economics, vol. 112, no. 1,

2010, pp. 1–24., doi:10.1111/j.1467-9442.2009.01592.

Roemer, John E. “Kantian Optimization: A Microfoundation for Cooperation.” Journal of Public

Economics, vol. 127, 2015, pp. 45–57., doi:10.1016/j.jpubeco.2014.03.011.

Roemer, John E. *How We Cooperate: A Theory of Kantian Optimization*. Yale University Press, 2019.

Rousseau, J.J. (2004). *The Social Contract*. Penguin Books.

Shinada, Mizuho, and Toshio Yamagishi. “Punishing Free-riders: Direct and Indirect Promotion of Cooperation. Evolution and Human Behavior, vol. 28, no.5, 330-339.

Singer, Peter. (1981). *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.

Street, S., 2006, “A Darwinian Dilemma for Realist Theories of Value,” Philosophical Studies, 127: 109–66.

Studtmann, Paul, and Shyam Gouri Suresh. “Universalizing and the We: Endogenous Game-Theoretic Deontology.” Economics & Philosophy, vol. 37, no. 2 July 2021, pp. 244 – 259., doi: https://doi.org/10.1017/S0266267120000279.

Tarnita, CE, T Antal, and MA Nowak. “Mutation-selection equilibrium in games with mixed strategies.” *Journal of Theoretical Biology* 261 (2009): *261*, 50-57.

Wiessner, P. “Norm enforcement among the Ju/’hoansi Bushmen.” *Hum Nat* **16,** 115–145 (2005). https://doi.org/10.1007/s12110-005-1000-9.

Zanella, Matteo, et al., “Dosage analysis of the 7q11.23 Williams region   
identifies BAZ1B as a major human gene patterning the modern human face and   
underlying self-domestication.” Science Advances 2019; 5.

1. In a related paper (currently in progress), we use simulations to show that the symmetric equilibrium strategy can be ES against both agents who always defect and those who always cooperate. Moreover, it is possible to demonstrate a phase transition along the dimension of the stakes of the interaction. When the stakes become large enough, autonomy is ES against both agents who always defect and those who always cooperate. [↑](#footnote-ref-1)