Diversity, tolerance and the social contract

Justin P. Bruner
University of California, Irvine, USA

Abstract

Philosophers and social scientists have recently turned to game theory and agent-based models to better understand the nature of conventions, norms, altruism and social contract formation. The stag hunt game is an idealization of social contract formation. Using the stag hunt game we attempt to determine what, if any, barrier diversity is to the formation of an efficient social contract. We uncover a deep connection between tolerance, diversity and the social contract. We investigate a simple model in which individuals possess salient traits and behave cooperatively when the difference between their trait and the trait of their counterpart is less than their "tolerance level." If traits are fixed and correspond to permanent or semi-permanent features of the individual such as religion or race social contract formation is a remote possibility. If traits are malleable social contract formation is possible but comes at the steep cost of diversity and tolerance – individuals are unwilling to cooperate with those much different from themselves. Yet homogeneity and intolerance are not a long term feature of the population. Overtime mutations allow for increasingly tolerant agents to prosper, thereby ushering in trait diversity. In the end, all reap the benefits of cooperation.

Keywords
Social Contract Theory, Game Theory, Diversity, Tolerance, Cultural Evolution, The Stag Hunt, Expanding Circle

1. Introduction

Does diversity hinder the formation of an efficient social contract? If individuals do not take into account their differences when deciding whether to cooperate, then the answer is "no". Unfortunately, this is rarely the case. It is not uncommon for humans and other organisms to condition their behavior on how similar they are to those with whom they interact. This tendency manifests itself in the choices made in strategic situations, such as social contract games.

Stag hunt games are idealizations of social contract formation. In recent years philosophers and social scientists have attempted to assess the plausibility of social contract formation by appealing to evolutionary game theory and agent-based modeling. According to much of the literature on the evolution of cooperation, different social mechanisms can make the formation of a social contract more or less feasible.[1] Specifically, Skyrms (2004) demonstrates that if individuals are embedded in a social network and employ an imitate-the-best update rule, stag hunters dominate the population. Zollman (2005) extends this work by allowing individuals to send costless pre-game signals to their neighbors before engaging in a stag hunt. Both Skyrms and Zollman demonstrate that the likelihood of a social contract dramatically increases when simple and realistic social mechanisms are taken into account.

So-called "similarity-based strategies" are another type of realistic social mechanism. Individuals possess a number of observable traits and condition their behavior in a game on how similar they are to their counterpart. In the social realm, similarity-based cooperation has been documented numerous times. Individuals in both natural and experimental settings seem to condition their behavior on how similar they are to their counterparts. Glaeser et al. (2000), for example, find little cooperation in trust games when the players are of different races. Strikingly, Krupp et al. (2008) find that individuals are more likely to contribute to a public good the more they physically resemble their fellow group members. These results are not limited to the laboratory. A number of natural experiments suggest that economic agents in real-life environments tend to employ similarity-based strategies. Miguel et al. (2005) document that regions in Africa characterized by high ethno-linguistic diversity tend to invest less in infrastructure and public goods.

The above examples seem to support a rather pessimistic story – cooperation seems possible but often comes at the price of diversity. Trust and aid are not extended to those who are too different from oneself. This leads to a natural question: is it possible to transform an intolerant group into an open and tolerant one? Peter Singer (1981) suggests it is possible for such a transition to naturally occur. Overtime, Singer contends, the circle of cooperation will slowly expand, thereby permitting increasingly distinct members to join in the cooperative enterprise. We observe a similar dynamic in this paper in section 6.

The scope of this paper is limited to investigating the effect similarity-based strategies have on the formation of a social contract. We will rely on a model loosely based on the collaborative work of Riolo et al. (2001).[2] In their model agents are endowed with a trait represented as a number in the interval from zero to one. Individuals cooperate if the distance between themselves and their partner in trait-space is less than their "tolerance level".

We'll find that in a number of situations, attaining the cooperative equilibrium in the stag hunt game is possible even though agents employ similarity-based strategies. If agents can change their traits with ease, cooperation is almost guaranteed. If traits are difficult to imitate because they correspond to permanent or semi-permanent features of the individual, such as race or culture, then social contract formation is improbable. In the case where traits are easily adopted, the population will naturally cluster in trait-space. These individuals will be rather intolerant and refuse to cooperate with those much different from themselves. Nonetheless, this clustering and intolerance allows for the formation of a social contract—all will be hunting stag. Furthermore, we'll see that this clustering in trait-space is not a permanent feature of the population. Due to mutations, individuals slowly become more tolerant and over time are willing to cooperate with increasingly larger areas of trait-space. Diversity and the social contract seem completely compatible.

2. The stag hunt

The stag hunt, shown in Table 1, is a game between two players. Each has the option of hunting stag or hare. If both individuals opt to hunt stag the operation is a success and they receive a large bounty. If one of them opts to hunt hare, the stag hunt fails and the lone stag hunter is left empty handed. Hunting hare results in a small reward, but said reward is not contingent on the counterpart's behavior. Hunting stag may result in a large reward but is inherently risky. When $S > H > V$ the two pure equilibria of the game are <stag, stag> and <hare, hare>.[3]

	Stag	Hare
Stag	S, S	V, H
Hare	H, V	H, H

Table 1: Normal form of the stag hunt with S> H > V.

This game is a stark contrast to the more popular prisoner's dilemma, shown in Table 2. While the stag hunt requires one to weigh the associated risks and benefits of hunting stag, it is always prudent to defect in the prisoner's dilemma. Defection is a strictly dominant strategy and hence is rational to perform regardless of the counterpart's behavior. Thus <defect, defect> is the sole equilibrium of the prisoner's dilemma.

Both the stag hunt and the prisoner's dilemma have many interpretations relevant to political and social philosophy. The stag hunt appears in Hume's *Treatise* in the form of a meadow-draining problem. Taking a modern example, James (2012) argues the stag hunt is representative of the situation we currently face with our global trade partners and hence is instructive when thinking of justice and collective-action on the international stage. For instance, countries morally motivated to adopt regulations that abate carbon emissions may fail to implement such policies without proper assurance others will follow suit (unilateral action is costly and does little to mitigate climate change). This situation is essentially a stag hunt; although all involved are eager to fulfill their moral duty and adopt pro-environment legislation, such activism is riskier than sticking with the status-quo.

Skyrms, echoing Hume and Rousseau, contends the stag hunt aptly models the strategic situation underlying the formation of a social contract.[4] For the state of nature to be difficult to transcend, it must be a Nash equilibrium. This corresponds to the stable but inefficient <hare, hare> equilibrium. Opting to form a social contract is inherently risky but is beneficial to all if realized. Analogously, hunting stag is inherently risky but pays great dividends when successful. Another line of literature, which can be traced to Rawls and Gauthier, takes the one-shot prisoner's dilemma to be the correct representation of the state of nature.[5] While it is compelling to view the prisoner' s dilemma as a social contract game, Skyrms (2004, 4-6) demonstrates that deciding how to behave in a repeated prisoner's dilemma is strategically identical to that of deciding how to behave in a one-shot stag hunt.[6] Thus, even if the interaction between agents in the state of nature is best modeled by the prisoner's dilemma, the repeated nature of these interactions directs our attention to the stag hunt.

	Cooperate	Defect
Cooperate	R, R	S, T
Defect	T, S	P, P

Table 2: Normal form of the prisoner's dilemma with T > R > P > S.

A recent project among philosophers and social scientists seeks to determine under what circumstances stag hunters flourish.[7] Unfortunately, the replicator dynamics rarely leads to the stag hunting equilibrium. Additionally, it has been shown that a stochastic evolutionary system spends the majority of its time at the hare hunting equilibrium. Surprisingly, we'll see that

allowing agents to employ similarity-based strategies greatly increases the probability of arriving at the cooperative equilibrium.

3. Similarity-based cooperation

Similarity-based cooperation is a deceptively simple concept. Agents condition their behavior in a game on a salient trait possessed by their counterpart. A trait could be anything from armpit scent to visual cues such as skin color, cultural emblems, or even hairdos. The agent cooperates if his counterpart's trait is sufficiently similar to his own. If the agents play the prisoner's dilemma, it is easy to see why similarity-based strategies will not result in sustained cooperation: a mutant possessing the correct trait defects and performs exceedingly well.

William Hamilton was the first in the biological literature to suggest the possibility of a similarity-based strategy.[8] Hamilton invites us to imagine a gene that regulates both the presence of an observable trait in the organism and the organism's propensity to cooperate with those possessing this trait. This so called "green-beard effect" is widely considered unfeasible.[9] While the possibility of hard-wired, similarity-based cooperation appears remote, theoretical biologists and social scientists have not abandoned the project. Many have attempted to show altruistic behavior can evolve with the help of similarity-based strategies.[10] For example, Riolo et al. posit a model with a continuum of potential traits. Individuals then determine whether to behave altruistically based on the Euclidean distance between themselves and their counterpart in trait-space.

One common finding is that cooperation in the prisoner's dilemma is possible but requires a cycling of traits. This so-called chromodynamics is due to the invasion of mutants. A group of cooperating agents all possessing green beards is invaded by a mutant green beard that defects when interacting with other green beards. Cooperation is still possible but requires that the agents condition their behavior on a new trait. Thus, green beards may be followed by purple beards which in turn are followed by blue beards, etc. The existence of cycling in nature seems doubtful because implicit in these theoretical models is the assumption that trait mutations occur much more rapidly than strategy mutations. In fact in some models it is assumed traits mutate on the order of two whole magnitudes faster than strategies mutate.[11] While it is questionable whether similarity-based strategies can sustain high levels of cooperation in the prisoner's dilemma, we'll see that it is much better suited to promote cooperative behavior in the stag hunt. It thus comes as little surprise that the strategic situation underlying one of the few documented cases of the green-beard effect in nature is reminiscent of the stag hunt.[12]

4. The model

This simple model is loosely based on work in Riolo et al. (2001). Individuals are assigned two values, a trait (also referred to as a tag) and a tolerance level, both represented by numbers drawn from a discrete uniform distribution from zero to one with 0.001 intervals. If individual *i* has a tolerance level of Tol_i, he will behave in the following manner when interacting with a second individual, *j*:

Stag if $|Tag_i - Tag_j| < Tol_i$

Hare if $|Tag_i - Tag_j| \geq Tol_i$

Note that an agent may be willing to cooperate with those outside of the [0, 1] interval of trait-space. A tolerance level of zero means one is an unconditional hare hunter, unwilling to

cooperate even with an individual endowed with identical traits. In Riolo et al. individuals with the same trait were artificially forced to cooperate with each other, and this strong assumption was in part responsible for their favorable results.[13]

We'll start with a population of 100 agents all with randomly sampled tags and tolerance levels. Each agent will be randomly paired with ten other individuals to play the stag hunt. Each interaction will result in a payoff.[14] We'll call the sum of these and only these ten payoffs the individual's "total payoff" (TP).[15] Agents will then be randomly paired once more for an imitation period. If an agent has a lower TP than her counterpart, she will adopt both the trait and tolerance level of the other player. If the two have the same TP, no imitation occurs.[16] We'll make slight alterations to the baseline model, and eventually introduce mutations to traits and tolerance levels.

5. No mutations

We begin by running 1,000 trials of the baseline model, described above. We find that without mutations, *all of these simulations result in universal stag hunting*—all random pairings result in both agents opting to hunt stag. What occurs is that the randomly distributed agents in trait-space begin to cluster together, making it easier for them to cooperate.[17] Clustering in trait-space is vital to establishing high levels of cooperation and is powerful enough to promote the formation of a social contract even when the agents themselves are not particularly tolerant.

Clustering occurs by chance. Two individuals relatively close to each other in trait-space happen to interact with one another. If they have sufficiently high tolerance levels, then they'll reap the benefits of stag hunting. These two agents typically outperform those with relatively higher levels of tolerance because the higher one's tolerance level, the more likely one is to be let down by one's partner—she'll go for the hare while you opt for stag. Thus those with higher tolerance levels will imitate one of the paired agents with lower tolerance. This results in a positive feedback loop. Before long the entire population is concentrated in a small interval of trait-space, all with sufficiently high levels of tolerance to cooperate with each other.

This finding is fairly robust, so much so that even including a substantial number of unconditional hare hunters (tolerance equal to zero) in the population does not prevent social contract formation. Table 3 shows that there are still a substantial number of simulations that converge to the cooperative equilibrium even in cases where half the population is initially hunting hare.

Number unconditional hare hunters	0	10	20	30	40	50	60	70	80
% of sims go to stag eq.	100	100	98	80	57	41	20	10	2

Table 3: Percentage of simulations that go to universal stag hunting as a function of the number of unconditional hare hunters in the initial population.

We now relax a vital assumption that seems responsible for our pleasant results. We have seen that if agents cluster in trait-space, cooperation is almost inevitable. However, this assumes that agents can easily change their traits. Relaxing this assumption may reduce the prospects of wide-scale cooperation. What does it mean for a trait not to be easily imitated? As put forth in the opening sections, traits are just observable features upon which the players condition their hunting behavior. This is an extremely broad definition and encompasses a number of different

features, such as clothing, cultural or religious symbols and even physiological attributes such as height, weight or race. Note that there is an important distinction between clothing and physiological features. Clothing, hairdos and shoes can all be manipulated much more easily than one's weight or race. We'd say weight is a *sticky* trait while clothing is a *plastic* trait: in the short span of an afternoon I can easily alter my wardrobe, but I cannot change my weight. Style and physiological features can be thought of as extreme cases. Accents are an ideal example of an intermediary case: accents are malleable but cannot usually be altered in the short term.

We'd expect more inflexible traits to inhibit clustering and thus lead to low levels of cooperation. To assess the effect trait plasticity has on cooperation, we run identical simulations as before, except with probability p the agent will adopt the trait of her counterpart, conditional of course on her counterpart having a higher TP. Thus with p=0.5 there is a 50 percent chance an agent will adopt the trait of her more successful imitation partner.[18] We run simulations with values of p = 0.2, 0.4, 0.6, 0.8 and 1.[19] See Figure 1.

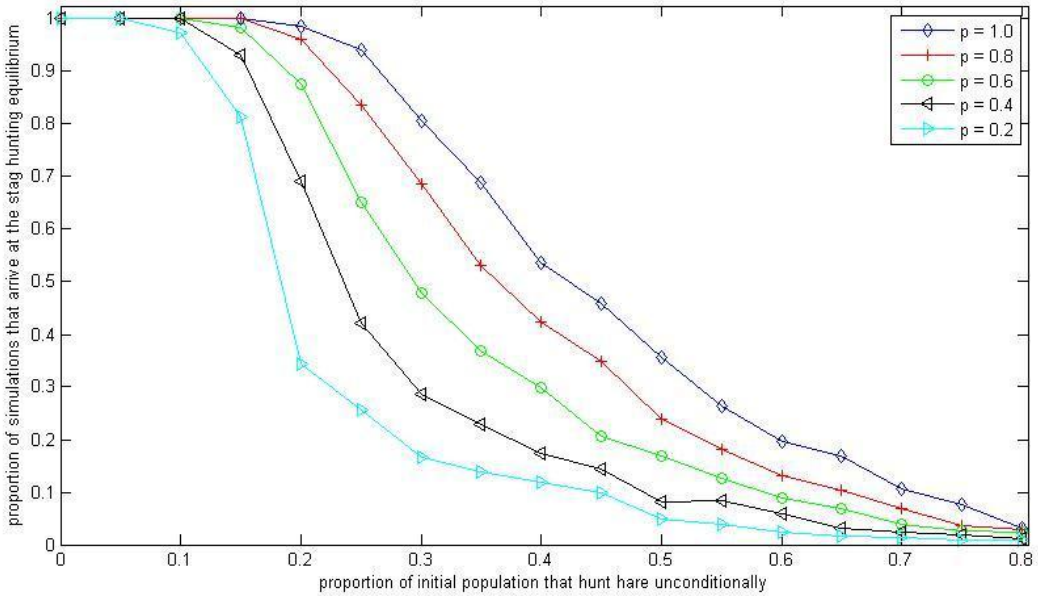

Figure 1: Percentage of simulations that arrive at the stag hunting equilibrium as a function of the number of unconditional hare hunters in the initial generation. Tag plasticity (p) varies from 0.2 to 1.

When there are no unconditional hare hunters in the initial population, trait plasticity is immaterial—all simulations result in universal stag hunting and sticky traits merely delay the inevitable convergence to the stag hunting equilibrium. If there are many unconditional hare hunters in the initial population, sticky traits result in low levels of cooperation. Since traits are sticky, two cooperating individuals may not immediately attract others to their location in trait-space, making the early stages of cluster formation extremely fragile.

6. Mutations

The previous section demonstrated the essential role intolerance and homogeneity have in the evolution of cooperation. Somewhat intolerant individuals with similar traits cooperate and over time draw a majority of the population to their region in trait-space. This clustering means stag hunting is just about guaranteed. Overall, the social contract is possible but comes at the price of diversity and results in the proliferation of rather intolerant agents only willing to cooperate with a thin slice of trait-space. This is not the end of the story. Clustering still occurs, but with reasonable mutation rates the lack of diversity and tolerance is transient.[20] In the long run, cooperation is possible and diversity can be preserved. We can have our cake and eat it, too.

In the prisoner's dilemma, mutations cause problems. Mutants with lower than average tolerance exploit the members of the cooperative cluster. A new cluster emerges elsewhere in trait-space and establishes high levels of cooperation until mutants once again invade this new group. This cat and mouse game is what Riolo et al. observed. Such cyclic behavior is not present when we consider the stag hunt. Once the population clusters in trait-space, intolerant mutants will not thrive. They will unnecessarily refuse to cooperate with fellow group members willing to hunt stag. As long as the number of such mutants is low, they will be outperformed by those with high tolerance.[21]

The expanding circle

To get a better sense of how mutations affect long-term behavior, we'll first start by examining the case in which the agents have already successfully clumped together and formed a social contract. Consider the dynamics of such a group of agents clustered in trait-space. We start with 100 individuals all with the same trait (0.5) and miniscule tolerance level (0.01). Additionally, with a 10 percent chance, a given agent's tolerance will be perturbed by a pull from N(0, 0.1). An agent's tag will for now not be perturbed. What occurs is the following: mutants with a tolerance level lower than the initial population average fare poorly. Due to their ultra-low tolerance level, they foolishly hunt hare with a community of individuals all willing to hunt stag. Unsurprisingly, mutants with a tolerance level equal to or greater than the initial 0.01 continue to successfully coordinate on stag hunting. *Thus there is a weak selection for higher tolerance levels*, so much so that after 500 generations the average tolerance level of the population has risen to an astounding 0.624.

We will now include mutations to traits to the above framework. We find that for a number of parameters this does not prevent the population from continuing to cooperate. Tolerance increases just as it did in the absence of trait mutations. What additionally occurs is a diffusion of agents throughout the trait-space. When the average tolerance of the cooperative cluster is low, any trait mutations will likely be selected against—mutants are too different to cooperate with. When the average tolerance of the population increases, mutants can thrive. Diversity is slowly regained as average tolerance increases.

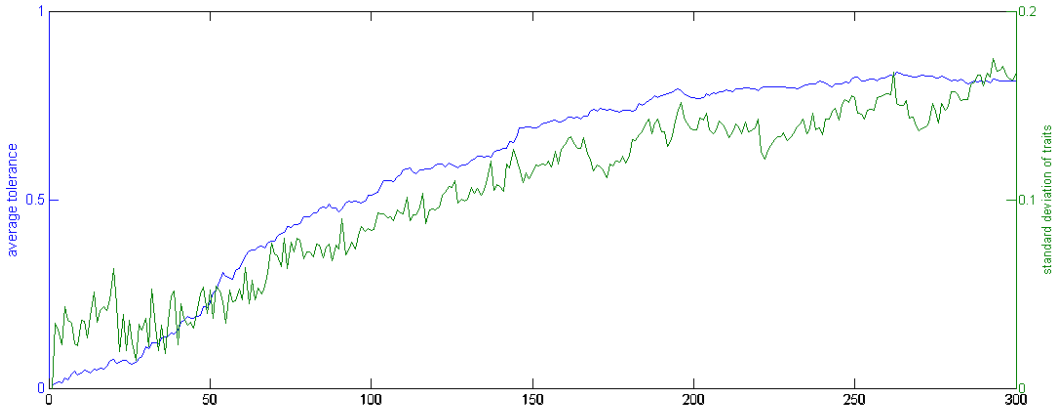

Figure 2: Average tolerance (blue) and standard deviation of the distribution of traits (green) over the course of the first 300 generations of a simulation. This data comes from one simulation consisting of 100 individuals all with an initial tolerance level of 0.001 and a trait values of 0.5.

This process is slow. Diversity is slowly regained because (i) perturbations to traits are small and (ii) traits that stray too far from the cooperative cluster fare poorly because they fall outside of the cluster's tolerance radius. Each generation has a slightly higher average tolerance level than previous generations, expanding the "radius" of cooperation. This circle expands slowly until agents from both extremes of trait-space can peacefully cooperate. In other words, with enough time, a highly intolerant and homogenous population can be transformed into a diverse and tolerant one.

When traits are sticky (i.e., people imitate traits less frequently than they imitate strategies) stag hunting can still thrive, but a lower trait mutation rate is necessary. If the trait mutation rate is high, then many agents accumulate on the periphery of the cooperative cluster, making it easier for those with a higher than average tolerance level to be let down by their counterparts, and thus allowing intolerant agents to flourish. If the trait mutation rate is low, cooperation is possible and in the long run the familiar dispersion of agents occurs throughout trait-space. Hence, sustaining a social contract is possible when traits are sticky, but hinges on the trait mutation rate.[22]

The contracting and then expanding circle

Let's put all of this together. We now start with individuals spread out randomly in trait-space and allow mutations to both tags and tolerance. As in Section V, we see a clustering in trait-space and all within the cluster are successfully hunting stag. Once this occurs we then observe a steady increase in tolerance accompanied by a slow spread of agents in trait-space. This two-stage process demonstrates that a successful social contract requires a certain amount of cohesion; however, this structure need not be permanent (See Figure 3). Once a social contract emerges, more and more of the trait-space can slowly participate.

The contracting-expanding dynamic is visually striking if we consider two dimensional trait-space. Agents still possess a tolerance level but now have two traits as opposed to just one. The distance between two agents is simply the Euclidean distance between points in two dimensional space. (See Figure 4).

Once again our results are less hopeful if traits are sticky. We run a hundred simulations in which traits are imitated with a one-tenth chance ($p = 0.1$) and find only 50 percent of these simulations result in universal stag hunting. Sticky traits affect both stages of the dynamic. Low trait plasticity makes it more difficult for clustering to occur. Additionally, if clustering is successful, too many mutants on the periphery of the cluster can destabilize the group and result in the population settling on the hare hunting equilibrium.[23]

7. Discussion

This paper investigates the deep connection between diversity, tolerance and the social contract. We found, in particular, that similarity-based strategies can promote cooperative behavior in many scenarios and, surprisingly, homogeneity and intolerance are often essential intermediate steps toward the formation of a social contract. However, as vital as homogeneity and intolerance are to cooperation, both are not long-term features of the population. Once the agents are cooperating, tolerant mutants will prosper and soon all of trait-space can participate in a thriving social contract. In the long run, a social contract amazingly does not come at the price of diversity.

These fortuitous results suggest a natural moral progression, namely, that intolerance and homogeneity often pave the way to diversity and wide-spread cooperation. Such a transition has been noted in modern times by Peter Singer. Singer observes a natural tendency for the 'circle of altruism' to "broaden from the family and tribe to the nation and race" and eventually go so far as to "extend to all human beings."[24] Figure 3 and Figure 4 nicely illustrate such a movement. Over time a social contract transforms from an exclusive enterprise to one that allows all to participate.

This is all with one wrinkle though. Singer takes this progression to be the result of reason, going as far to suggest the expanding circle is either an "accident of history […] or the direction in which our capacity to reason leads us." (113) This appears to be a false dichotomy. The results uncovered in this paper are neither accidental nor are they driven by deliberate moral thought and reason. Instead, the border of the circle expands due to simple agents merely imitating those who are successful. The continual stream of mutations and experimentation is enough to cause boundedly rational individuals to become more tolerant. No moral reflection or reasoning is necessary.

While our results are promising, we should remember the rather intuitive finding that the social contract is less likely when traits are sticky. In the presence of persistent diversity, establishing an efficient social contract seems improbable. Further exploration of sticky traits seems fruitful. For instance, when individuals have multiple observable traits can agents learn to only take into account those traits that are plastic? If so, cooperation seems unavoidable. Imagine a population of agents each with two traits, one plastic and the other permanent. Individuals may measure their distance in trait-space by calculating the Euclidean distance between them in two dimensional space. However, agents could just as easily only base their decision on how closely they resemble their counterpart with respect to one trait. For example, individuals may decide to base their behavior solely on how closely they resemble their counterpart with respect to the plastic trait. How would a population of racists (those conditioning their behavior on fixed traits alone) and open-minded individuals (those conditioning their behavior on plastic traits alone) fare? This is an open question.

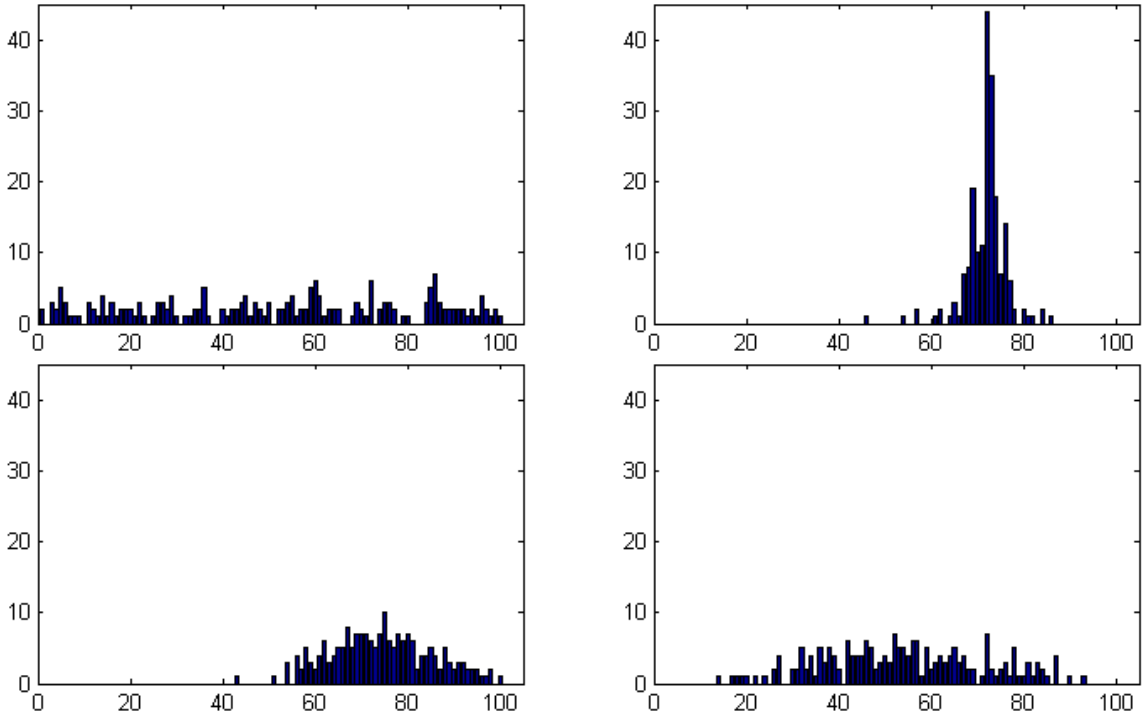

Figure 3: Distribution of traits. T = 1 (top left), T = 150 (top right), T = 1000 (bottom left), T = 2000 (bottom right). Initial tolerance was drawn from the distribution U[0, 0.1].

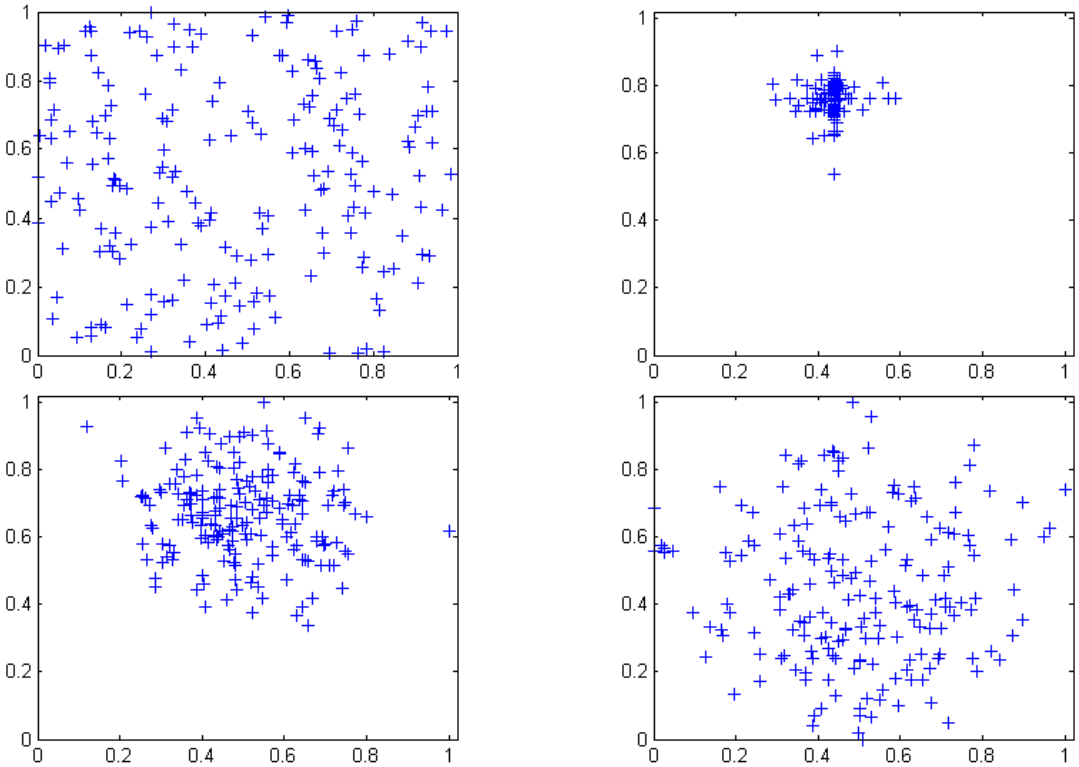

Figure 4: Two trait case. Distribution of traits in 2-dimensional trait-space. T = 1 (top left), T = 40 (top right), T = 1000 (bottom left), T = 2000 (bottom right). Initial tolerance was drawn from distribution U[0, 0.1].

Acknowledgements

I'd like to thank Brian Skyrms, Elliott Wagner, Simon Huttegger and Greg McWhirter. This paper benefited from the feedback of participants at the Social Dynamics Seminar at UC Irvine, the IMBS graduate student conference at UC Irvine, Philosophy of Biology in the Desert at Arizona State University and Formal Ethics 2012 hosted by the Center for Mathematical Philosophy, LMU, Munich.

References

Antal T, Ohtsuki H, Wakeley J, Taylor P and Nowak M (2009) Evolution of cooperation by phenotypic similarity. *Proc. Natl. Acad. Sci. USA* 106: 8597-8600.

Alexander JM (2007) *The Structural Evolution of Morality*. Cambridge University Press.

Axelrod R (1984) *The Evolution of Cooperation*. NY: Basic Books.

Axelrod R, Hammond RA and Grafen A (2004) Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution* 58: 1833-1838.

Axtell R, Epstein J and Young P (2006) The emergence of class in a multi-agent bargaining model. In *Generative Social Sciences: Studies in Agent-Based Computational Modeling*, ed. Joshua Epstein. Princeton: Princeton University Press.

Dawkins R (1987) *The Extended Phenotype*. Oxford University Press.

Gauthier D (1969) *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford University Press.

Glaeser E, Laibson D, Scheinkman J and Soutter C (2000) Measuring trust. *The Quarterly Journal of Economics* 155: 811-846.

Hales D (2005) Change your tags fast! A necessary condition for cooperation? *Multi-Agent and Multi-Agent Based Simulations* 3415: 89-98.

Hamilton WD (1964) The genetical evolution of social behavior II. *Journal of Theoretical Biology* 7:17-52.

Hobbes T (1994) *Leviathan.* Ed. E Curley, Indianapolis: Hackett Publishing Co.

Hume D (2000) *A Treatise of Human Nature*. Oxford University Press, eds. D F Norton and M. J. Norton edition.

James A (2012) *Fairness in Practice: A social contract for a global economy*. Oxford University Press.

Jansen V and Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440: 663-666.

Krupp D, Debruine L, and Barclay P (2007) A cue of kinship promotes cooperation for the public good. *Evolution and Human Behavior* 29: 49-55.

Miguel E and Gugerty M (2005) Ethnic diversity, social sanctions and public goods in Kenya. *Journal of Public Economics* 89: 2325-2368.

Moehler M (2009) Why Hobbes' state of nature is best modeled by an assurance game. *Utilitas* 21: 297-326.

Nowak M (2006) Five rules for the evolution of cooperation. *Science*, 314: 1560-1563.

Pollack G (1989) Evolutionary stability of reciprocity in a viscous lattice. *Social Networks* 3: 175-212.

Queller D, Ponte E, Bozzaro S and Strassmann J (2003) Single-gene greenbeard effects in the social amoeba dictyostelium discoideum. *Science* 299: 105-106.

Rawls J (1971) *A Theory of Justice*. Cambridge: Harvard University Press.

Riolo R, Cohen M and Axelrod R (2001) The evolution of cooperation without reciprocity. *Nature* 414: 441-443.

Roberts G and Sherratt T (2002) Does similarity breed cooperation? *Nature* 418: 499-500.

Rousseau J (1755) *A Discourse on Inequality*. Penguin Books, Trans. M. Cranston (1984) edition.

Singer P (1981) The Expanding Circle: Ethics and Sociobiology. Princeton University Press.

Skyrms B (2002) Signals, evolution and the explanatory power of transient information. *Philosophy of Science* 69: 407-428.
Skyrms B (2004) *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.
Skyrms B and Zollman K (2010) Evolutionary considerations in the framing of social norms. *Politics, Philosophy & Economics* 9: 265-273.
Smead R and Huttegger S (2011) Efficient social contracts and group selection. *Biology and Philosophy* 26: 517-531.
Wagner E (2012) Evolving to divide the fruits of cooperation. *Philosophy of Science* 79: 81-94.
Weibull J (1995) *Evolutionary Game Theory*. Cambridge: MIT Press.
Zollman K (2005) Talking to neighbors: the evolution of regional meaning. *Philosophy of Science* 72: 69-85.

---

[1] For prime examples, see Axelrod (1984), Pollack (1989). For an in-depth but partial overview of the evolution of cooperation research, see Nowak (2006).

[2] There are a number of theoretical models that attempt to demonstrate similarity-based strategies lead to high levels of cooperation in the prisoner's dilemma. While many of these findings appear promising, positive results often hinge on a few rather extreme assumptions implicit in the underlying models. We'll see that cooperation in the stag hunt does not require such strict assumptions. See Gilbert Roberts and Thomas Sherratt (2002) for a detailed criticism of Riolo et al.

[3] In game-theoretic terms the <stag, stag> equilibrium is pareto dominant and the <hare, hare> equilibrium is considered risk dominant when $H > .5S > 0$. There also exists an unstable mixed Nash equilibrium.

[4] This sentiment is shared by Edwin Curley in the introduction to *Leviathan* (1994). Other arguments in favor of the stag hunt can be found in Michael Moehler (2009).

[5] In A Theory of Justice Rawls claims ''Hobbes's state of nature is the classical example" of the prisoner's dilemma.

[6] This assumes agents only have the choice of two strategies: "always defect" or the grim trigger strategy (cooperate until your counterpart defects and then continue to defect for the remainder of the interaction).

[7] Examples abound. For a few, see Alexander (2007), Wagner (2012), Smead and Huttegger (2010).

[8] Robson (1990) was the first to formally demonstrate how costless pre-game signals can foster high levels of cooperation in the prisoner's dilemma, albeit for brief periods of time.

[9] Richard Dawkins, in the *Extended Phenotype*, both coins the term "green beard" and argues against the possibility of such similarity-based strategies.

[10] For example see Antal et al. (2009), Axelrod et al. (2004) and Jansen and Baalen (2006).

[11] David Hales points this fact out. Additionally, Hales demonstrates through the use of simulation that cooperation is highly unlikely when traits and strategies mutate at the same rate.

[12] Specifically, the amoeba Dictyostelium discoideum possesses a "green beard" gene. When food is scarce the single-celled organisms huddle together and congeal into a mass perched upon a stem. If they are able to successfully coordinate, the majority of them survive; but if they fail to aggregate, the individual amoeba will perish. The gene csA encodes for the cell adhesion gp80 protein. Those without this protein are left behind when the aggregation process begins. The underlying strategic interaction here resembles a stag hunt. Establishing a protective sanctuary is

possible but requires large-scale cooperation. Going it alone minimizes uncertainty but leads to a sub-optimal result. See Queller (2003) for more details.

[13] Sherratt and Roberts provide a detailed criticism of a few key assumptions in the Riolo et al. model. Relaxing these assumptions led to low levels of cooperation in the prisoner's dilemma, but as we will see, does not have the same effect when the stag hunt is considered.

[14] The payoffs we will use are S = 3, H =2. As the ratio of S to H increases the basin of attraction for the stag hunting equilibrium expands. For example, if we use the payoff S = 15, H = 7 our results greatly improve. These payoffs are taken from Brian Skyrms (2002).

[15] It may be the case the agent plays the stag hunt more than ten times, but only these ten payoffs will count toward her TP.

[16] Models of imitation are often formally equivalent to the more traditional replicator dynamic. See Jorgen Weibull (1995) for details.

[17] This is the case even if we limit the initial tolerance levels by drawing instead from a more restricted discrete uniform distribution of zero to 0.1. By the end of the simulation, agents have very low levels of tolerance, but stag hunting is still possible because the population clusters up tightly in trait-space. This cluster is stable because unlike the prisoner's dilemma, once all are cooperating it no longer pays to have low tolerance.

[18] She will still always adopt the tolerance level of the imitation partner who outperformed her.

[19] Surprisingly, little research has been done on this topic. Axtell et al. (2006) investigate a bargaining model in which agents are endowed with permanent traits (p=0). Additionally, Skyrms and Zollman (2010) examine a model of agents with permanent traits (p = 0) playing the hawk-dove game. I am not aware of any work that investigates intermediary cases where p is between zero and one.

[20] Since our model is intended to explicitly model cultural evolution, we interpret mutations as either experimentation on the part of the agent (they make a deliberate choice to try a new strategy) or a failure on their part to correctly implement the strategy they choose (a trembling hand, as it is known in the economics literature).

[21] For our purposes we will model mutations in the following fashion: traits and tolerance levels are perturbed by a draw from a normal distribution with mean zero and a standard deviation of 0.1. We allow tolerance levels to go below zero and restrict our tags to the interval [0, 1]. Riolo et al. modeled mutations in a similar fashion, except tags were replaced by a draw from the uniform distribution U[0,1]. This alteration will not significantly change the qualitative results of our simulation.

[22] The less sticky a trait is, the less trait mutation rates matter. When traits are completely plastic we stay at the stag hunting equilibrium when the trait mutation rate is low (0.01) as well as high (0.15). Of course, if trait mutations are too frequent (0.5) then clustering is no longer possible and we revert to the hare hunting equilibrium.

[23] However, if we reduce the frequency of trait mutations, our results change for the better. Nearly all simulations arrive at the stag hunting equilibrium when the frequency of trait mutations is one-tenth that of the frequency of tolerance mutations. It seems sensible to assume that since traits themselves are sticky, mutations to them occur relatively infrequently.

[24] Peter Singer, *The Expanding Circle*, page 120. While Singer and I seem to slightly differ in our use of the word cooperation (Singer is concerned with altruism – costly acts which help others – while our concern has been primarily with collective action) I find it immensely striking

that the general ‘moral’ we both come to remains the same. Namely, there is a tendency for individuals to naturally cooperate with less and less restrictive subsets of the population.