Article

Politics, Philosophy & Economics
1–20

DOI: 10.1177/1470594X14560763
ppe.sagepub.com
SAGE

Diversity, tolerance, and the social contract

Justin P Bruner
Australian National University, Australia

Abstract

Philosophers and social scientists have recently turned to game theory and agent-based models to better understand 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 semipermanent 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, that is, individuals are unwilling to cooperate with those much different from themselves. Yet homogeneity and intolerance are not a long-term feature of the population. Over time 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

Corresponding author:
Justin P Bruner, School of Philosophy, Australian National University, 4105 Coombs Building, Australia.
Email: justin.bruner@anu.edu.au

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 under what conditions social contract formation is possible 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, a social contract can be secured. Zollman (2005) extends this work by allowing individuals to send costless pregame 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.

The 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. (2007) 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 and Gugerty (2005) document that regions in Africa characterized by high ethnolinguistic 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. Over time, 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 article in the section on Mutations.

Interest in tolerance is not restricted to moral and political philosophers alone. Biologists and social scientists alike seek to better understand the means by which tolerant behavior can be brought about. Milton Friedman (1962), for example, in *Capitalism and Freedom* argues that a competitive exchange economy will over time weed out discriminatory practices in the business community. Companies that take into account the

attributes of others that are impertinent from an economic point of view will be slowly driven out of business by competitors who only base their decision on economic values. In Friedman's view, there are strong self-interested reasons that compel people to become more tolerant. Friedman's narrative is in line with the findings of this article. As we'll soon see, once a social contract is successfully established individuals have self-interested reasons to slowly become increasingly more tolerant.

Biologists and biological anthropologists have also realized the importance of uncovering mechanisms that promote tolerance. Unfortunately, most of the work by social scientists and theoretical biologists has instead been preoccupied with the task of *explaining the emergence* of ethnocentric behavior, not exploring means of reducing it,[2] with a few exceptions.[3] Smith and Szathmary (2000) note that while ethnocentric behavior has the upshot of promoting group cohesion and high levels of cooperation *within* communities, it nonetheless leads to ethnic conflicts and violence between groups. While they do not offer any concrete recommendations, they encourage others to study different means of reducing the levels of ethnocentrism.

The scope of this article 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).[4] In their model agents are endowed with a trait represented as a number in the interval from 0 to 1. Individuals cooperate if the distance between themselves and their partner in trait space is less than their 'tolerance level'.

We follow the lead of Brian Skyrms and take the stag hunt game to capture the strategic considerations individuals face when determining whether to form or revise a social contract (more about this in the subsequent section). Like both Skyrms and Kenneth Binmore, we are primarily concerned with how individuals in a community can coordinate their actions with one another. A stable arrangement that allows individuals to successfully coordinate with others will be referred to as a social contract. We, for simplicity, restrict our attention to cases where there are only dyadic interactions – that is, social contract games are two-person games. While none of the major figures in the social contract tradition conceived of social contracts as attaining between just two individuals, many modern game-theoretic approaches to the social contract make this idealizing assumption. Binmore (2005), for example, models the social contract as a two-person bargaining game.[5] If we conceive of a social contract as a stable arrangement that facilitates the coordination of individual actions in a society (as Skyrms and Binmore do), then investigating two-person games is a reasonable starting place.

We'll find that in a number of situations, attaining the pareto-dominant 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 semipermanent 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

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

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

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

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

experimentation, 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.

The stag hunt

The stag hunt, shown in Table 1, is a game between two players (Rousseau, 1755). 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 >.[6]

This game is a stark contrast to the more popular prisoner's dilemma, as 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 (2000) *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.[7]

Skyrms, echoing Hume and Rousseau, contends that the stag hunt aptly models the strategic situation underlying the formation of a social contract.[8] 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.[9] Another line of literature, which can be traced to Rawls (1971) and Gauthier (1969), takes the one-shot prisoner's dilemma to be the correct representation of the state of nature.[10] While it is compelling to view the prisoner's dilemma as a social contract game, Skyrms (2004: 4–6) demonstrates that when the strategy choice is between 'grim trigger' and 'always defect', 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.[11] Thus, there is a sense in which 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.[12]

A recent project among philosophers and social scientists seeks to determine under what circumstances stag hunters flourish.[13] 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.

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 her counterpart's trait is sufficiently similar to her own. If the agents play the prisoner's dilemma, it is easy to see why similarity-based strategies will not result in sustained cooperation, that is, a mutant possessing the correct trait defects and performs exceedingly well.

William Hamilton (1964) was the first in the biological literature to suggest the possibility of a similarity-based strategy.[14] 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.[15] While the possibility of hardwired, similarity-based cooperation appears remote, theoretical biologists and social scientists have not abandoned the project. Many have attempted to show that altruistic behavior can evolve with the help of similarity-based strategies.[16] For example, Riolo et al. (2001) 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, and so on. 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.[17] 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.[18]

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 0 to 1 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:

$$\text{Stag if } \left|Tag_i - Tag_j\right| < Tol_i$$
$$\text{Hare if } \left|Tag_i - Tag_j\right| \geq Tol_i$$

A tolerance level of 0 means one is an unconditional hare hunter, unwilling to cooperate even with an individual endowed with identical traits.[19] In Riolo et al.'s study, 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.[20]

We'll start with a population of 100 agents all with randomly sampled tags and tolerance levels. Each agent will be randomly paired with 10 other individuals to play the stag hunt. Each interaction will result in a payoff.[21] We'll call the sum of these and only these 10 payoffs the individual's 'total payoff' (TP).[22] 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.[23] We'll make slight alterations to the baseline model and eventually introduce mutations to traits and tolerance levels.

It is worth being explicitly clear at this point that while many of the articles cited in the section on Similarity-based cooperation pertain to biological evolution, the model explored in this article is a model of cultural evolution. The agents in the model are presumed to be boundedly rational humans, striving to maximize their payoffs by imitating those more successful than themselves. While imitation makes perfect sense with respect to the publically advertised 'tags', it is less clear how the imitation process works when it comes to tolerance levels. Tolerance levels are private traits, meaning one cannot directly observe the tolerance level of another. Yet one can attempt to infer the tolerance level of another by observing them over the course of many interactions. We modify the baseline model so the imitation of tolerance levels occurs with some noise– that is, when imitating another, the tag is perfectly imitated but the tolerance level is imperfectly imitated. Incorporating noise into the imitation process does not appear to affect our

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.

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

results.[24] Yet as the size of the group grows, it is implausible to assume that each individual tracks the past behavior of each and every other group member. So how can individuals make an educated guess as to the tolerance level of their imitation partner? It is common in these game-theoretic models to assume the existence of some sort of social mechanism that can provide agents with detailed information about others in the population. For example, individuals could attain a rough estimate of the tolerance level of another through the help of a gossip network, responsible for keeping track of others' past behavior. Such networks are frequently assumed but not explicitly modeled in the evolution of cooperation literature.[25] Even a somewhat unreliable and noisy gossip network would nonetheless enable agents to acquire a rough estimate of the tolerance level of others in the population.

No mutations

We begin by running 1000 trials of the baseline model described previously. 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 begin to cluster together in trait space, making it easier for them to cooperate.[26] 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 will go for hare while you opt for stag. Thus those with higher tolerance levels will imitate one of the paired agents with lower tolerance, adopting both their trait and their tolerance level. The more populated this region of trait space is, the more likely individuals from this region will interact with each other and successfully coordinate on stag hunting. This results in a positive feedback loop and before long the entire population is concentrated in this small interval of trait space, all with sufficiently high levels of tolerance to cooperate with each other.

This finding is robust, so much so that even including a substantial number of unconditional hare hunters (tolerance equal to 0) 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 an unconditional hare hunter.

We now relax a key assumption that seems responsible for our pleasant results. We have seen that if agents cluster in trait space, cooperation is almost inevitable. However,

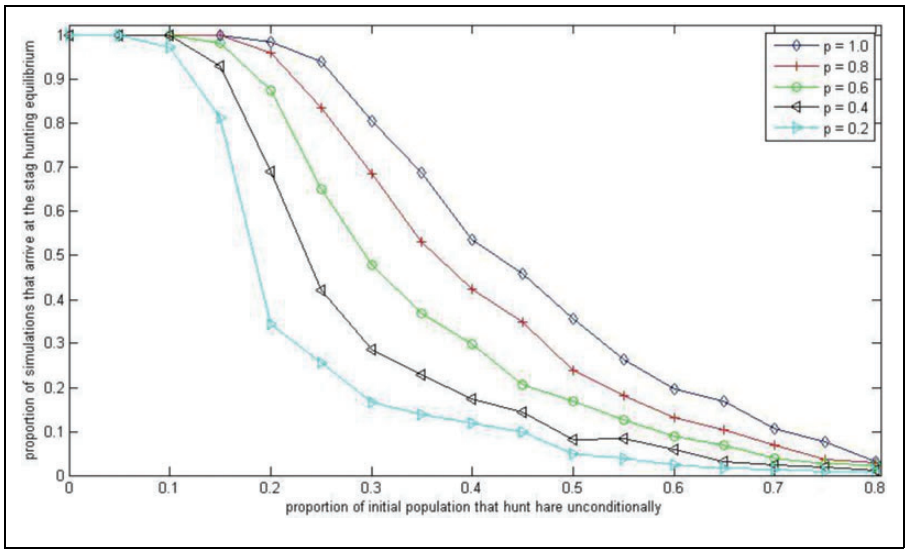

Figure 1. Percentage of simulations that arrive at the stag hunting equilibrium as a function of the number of unconditional hare hunters in the initial generation. Tag plasticity (p) varies from 0.2 to 1.

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 race is a *sticky* trait, while clothing is a *plastic* trait, that is, in the short span of an afternoon one can easily alter one's wardrobe, but one can never change race. 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.[27] We run simulations with values of $p = 0.2$, 0.4, 0.6, 0.8, and 1,[28] see Figure 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.[29] Since traits are sticky, two cooperating individuals may not immediately attract others to their location in trait space. In general, clustering has to occur quickly because as more time passes it becomes more likely that the founding members of the cluster imitate either the tolerance or trait of an individual not in the cluster. Thus, sticky traits make the early stages of cluster formation extremely fragile.

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. However, this is not the end of the story. With reasonable mutation rates clustering can still occur, but the lack of diversity and tolerance is transient.[30] 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.[31]

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 ultralow 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

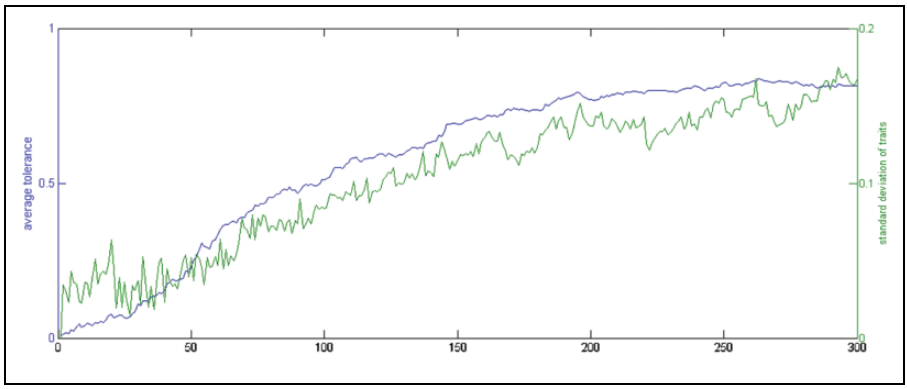

Figure 2. Average tolerance (blue) and standard deviation of the distribution of traits (green) over the course of the first 300 generations of a simulation. These data come from one simulation consisting of 100 individuals, all with an initial tolerance level of 0.001 and a trait value of 0.5.

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).

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 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, thus allowing intolerant agents to flourish. This then pushes the population to the hare hunting equilibrium. If the trait mutation rate is low, cooperation is possible and in the long run the familiar dispersion of agents throughout trait space occurs. Hence, sustaining a social contract is possible when traits are sticky but hinges on the trait mutation rate.[32]

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 the section on No mutations, 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

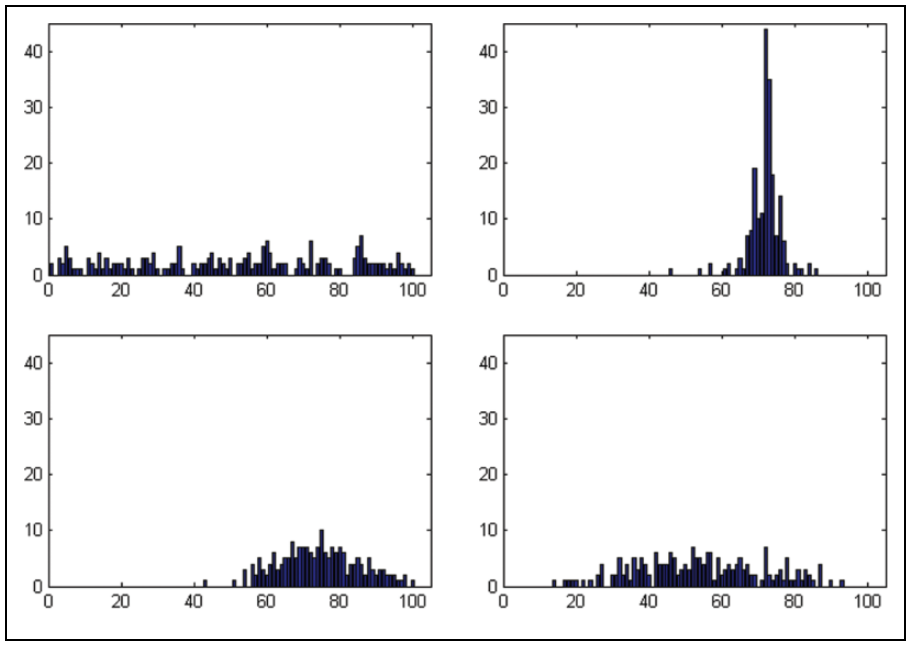

Figure 3. Distribution of traits. $T = 1$ (top left), $T = 150$ (top right), $T = 1000$ (bottom left), and $T = 2000$ (bottom right). Initial tolerance was drawn from the distribution $U[0, 0.1]$ and $p = 1$.

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 0.15 chance ($p = 0.15$) and find only 59 percent of these simulations result in universal stag hunting. As it becomes increasingly difficult to imitate the traits of others, we are less likely to observe this contracting–expanding dynamic. The percentage of simulations that underwent the contracting–expanding dynamic is 2, 13, 59, 76, 87, 91, and 100, for a p of 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, and 0.4, respectively.[33] 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.[34]

Discussion

This article investigates the deep connection between diversity, tolerance, and the social contract. We found, in particular, that similarity-based strategies can promote

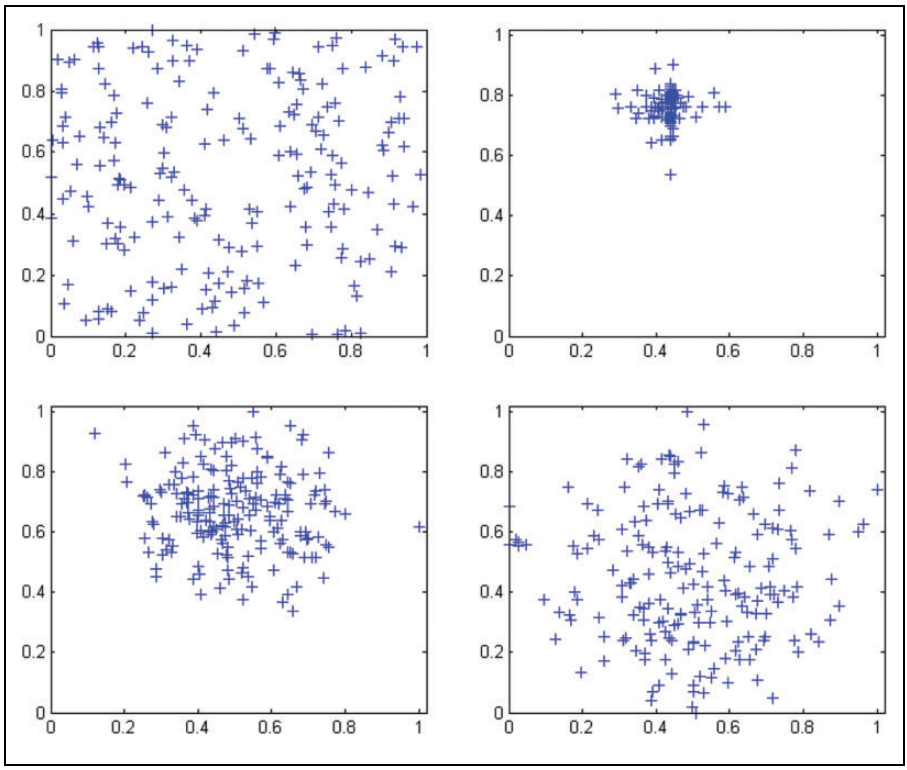

Figure 4. Two trait case. Distribution of traits in TWO-dimensional trait space. $T = 1$ (top left), $T = 40$ (top right), $T = 1000$ (bottom left), and $T = 2000$ (bottom right). Initial tolerance was drawn from distribution $U[0, 0.1]$ and $p = 1$.

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 widespread 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'.[35] Figures 3 and 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'. (Singer, 1981: 113) This

Table 4. A modified stag hunt game.

	A	B	C
A	4, 4	0, 3	0, 2
B	3, 0	3, 3	0, 2
C	2, 0	2, 0	2, 2

appears to be a false dichotomy. The results uncovered in this article are neither accidental nor 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.

Thus our results seem more in line with Milton Friedman's than Singer's. Once a social contract is established, it is in the agent's interest to have a higher tolerance level. Yet there is an important difference that distinguishes our account from Friedman's. On Friedman's account, one always has self-interested reasons to be more tolerant. This is not true in our model. Too much change too quickly will be selected against. Consider the situation in which the population is crammed into a region of trait space and for the most part all have the minimal levels of tolerance that guarantee near universal stag hunting. If an agent increases her tolerance too much, then she'll cooperate with those outside the cooperative circle unwilling to cooperate with her. Ergo, she would have benefited from being less tolerant. Increasing one's tolerance level by small increments is individually rational, but becoming an unconditional stag hunter may not be. As a result, the transition from an intolerant society to a tolerant one is a slow and gradual process – too much change in a short time period will be selected against.[36]

It is also worth mentioning that so far we've been implicitly assuming that there is just one viable social contract. This of course is not the case. In reality, there are a plethora of contracts and there exist real differences between the various feasible social contracts. Some, for example, are more efficient than others. How can we transcend an inefficient social contract? How is reform possible? In many ways, the problem of social contract *reform* is very similar to the problem of social contract *formation*. As a result, the solution we propose to the problem of reform will be very similar to the findings we've uncovered in this article. Traits and similarity-based strategies help facilitate institutional change. Consider a slightly modified stag hunt game with not two but three equilibrium. The game is almost identical to the stag hunt game, except that we allow for a third strategy.[37]

Consider that initially individuals only had strategies B and C available to them. In other words, they were playing the game from Table 4 instead of Table 1. We will continue to assume that agents possess a trait and a tolerance level, and play B with those in their tolerance radius and C with those outside the radius. By the end of these simulations we'd end, as we did in the section on No mutations, with all individuals spread out in trait space, and all with a sufficiently high tolerance level to play B with any other member in the population. Now once this (B, B) social contract is established, let's introduce strategy A (perhaps strategy A was not a viable strategy before). How do we now make the

transition from the (B, B) equilibrium to the (A, A) equilibrium? Consider the following series of events. By chance, two agents very close to each other in trait space experiment and each adopt a strategy that allows them to play A against those close to them in trait space and B against anyone outside of this 'A tolerance radius'. Just as in the section on No mutations, these two individuals will do slightly better than the rest of the population. This will in turn attract more of the population to this area of trait space, meaning even more agents will be willing to play A, as opposed to B. In the end, we will recover the familiar contraction–expansion dynamic identified in the section on Mutations – the final result being a transition from the (B, B) equilibrium to the (A, A) equilibrium.

In this way, traits allow those yearning for change to successfully interact with like-minded individuals while simultaneously ensuring reformers do not waste their time playing A with those unwilling to deviate from the status quo. This is fascinating for it demonstrates there is a natural and straightforward way in which reform is possible. Even more surprising, *this explanation does not rely on group selection*. Binmore (2005), for example, has argued informally that group selection is responsible for ensuring that efficient social contracts are adopted by societies.[38] We should expect evolution to 'succeed in selecting one of the efficient equilibrium [because] societies operating [with an efficient social contract] will grow faster' (p. 12). If we assume 'societies cope with population growth by splitting off colonies which inherit the social contract of the parent society', efficient social contracts should abound (p. 12). Traits and similarity-based strategies can successfully allow for the transition from inefficient social contracts to efficient ones. Group selection may speed up this process, but it is by no means required.

Finally, while the results of this article 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.

Acknowledgements

I would like to thank Brian Skyrms, Simon Huttegger, Hannah Rubin, Gregory McWhirter and especially Elliott Wagner for helpful comments on previous versions of this paper. I would also like to thank the participants of the Social Dynamics seminar at UC Irvine, the IMBS Graduate Student Conference 2012 at UC Irvine, Philosophy of Biology in the Desert 2011 at Arizona State University and Formal Ethics 2012 at the Munich Center for Mathematical Philosophy.

Notes

1. For prime examples, see Axelrod (1984) and Pollack (1989). For an in-depth but partial overview of the evolution of cooperation research, see Nowak (2006).
2. See, for example, Axelrod and Hammond (2006).
3. Muldoon et al. (2012) and Grim et al. (2005).
4. 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. (2001)
5. For further examples see Skyrms (1996) and of course Skyrms (2004).
6. In game-theoretic terms the <stag, stag> equilibrium is pareto-dominant and the <hare, hare> equilibrium is considered risk dominant when $H > 0.5$ $S > 0$. There also exists an unstable mixed Nash equilibrium.
7. This is of course a complex example, and many pertinent details that would complicate the analysis are left out (for example, there may be disagreement among these morally motivated individuals as to what course of action is best. See Kavka (1995) for more about the dilemmas morally perfect people may find themselves in).
8. 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). Additionally, although Kavka (1986) is often taken as arguing that the prisoner's dilemma is the best representation of Hobbes' (1994) state of nature, there is evidence that he believed the stag hunt was a more apt representation. On pages 113–116, Kavka introduces a new game he refers to as the *N*-person quasi-prisoner's dilemma and argues that if we are being very careful, it is technically *this* game that individuals in the state of nature play. Although it appears he failed to notice it, this *quasi-prisoner's dilemma is just the N-person stag hunt*.
9. Binmore (1994) has a slightly different interpretation of the stag hunt game. He conceives of the stag hunt as capturing the strategic situation we face when determining whether to go through with reform or not. In this case the 'state of nature' equilibrium is interpreted as the status quo, and the 'social contract' equilibrium is interpreted as the new and improved state of affairs. Thus the results of this article can be thought of as shedding light on both (i) the question of when a group can transcend the state of nature and (ii) the question of when reform is possible. For the rest of the article, we'll refer to the stag hunt game as a social contract formation game, although we believe Binmore's interpretation equally applies.
10. In *A Theory of Justice*, Rawls (1971) claims Hobbes's state of nature is the classical example of the prisoner's dilemma.
11. Grim trigger commands agents to cooperate until their counterpart defects, in which case one plays defect for the remainder of the interaction. The grim trigger is not the only strategy that converts the repeated prisoner's dilemma to the one-shot stag hunt. For example, the popular tit-for-tat (imitate the act performed by your counterpart in the last round of the game) can also convert the repeated prisoner's dilemma to a one-shot stag hunt when the only other available option is always defect.
12. One may wonder if it makes any sense to talk of repeated interactions in the Hobbesian state of nature (I thank an anonymous referee for bringing this criticism to my attention). If I cooperate

while my counterpart defects, I am dead and the game screeches to a halt. Death may be one outcome, but I believe it is premature to assume that this is the only outcome or even the most likely outcome of confrontation in the state of nature. Kavka (1986), for example, suggests that one may attack or 'anticipate' (that is, defect) against another in the state of nature for the express purpose of physically dominating one's opponent and thereby forcing him to protect you against others. Vanderschraaf (2006) argues that it is implausible that one could master another in the state of nature for more than a brief time, instead interpreting the defect or 'anticipate' strategy as the 'seizure of some of their power. Often one anticipates against another by seizing some of the goods she possesses. One can also anticipate against another by occupying some of her physical and mental powers in a conflict' (p. 247).

13. Examples abound. For a few, see Alexander (2007), Wagner (2012), Smead and Huttegger (2011).
14. 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.
15. Richard Dawkins (1987), in the *Extended Phenotype*, both coins the term 'green beard' and argues against the possibility of such similarity-based strategies.
16. For example, see Antal et al. (2009), Axelrod et al. (2004), and Jansen and Baalen (2006).
17. David Hales (2005) 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.
18. 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 suboptimal result. See Queller et al. (2003) for more details.
19. Note that an agent may also be willing to cooperate with those outside of the [0, 1] interval of trait space.
20. 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.
21. 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).
22. It may be the case the agent plays the stag hunt more than 10 times, but only these 10 payoffs will count toward her total payoff.
23. Models of imitation are often formally equivalent to the more traditional replicator dynamic. See Jorgen Weibull (1995) for details.
24. We alter the baseline model in the following way: whenever an agent imitates the tolerance level of another, her new tolerance level is the tolerance level of her imitation partner plus a draw from the normal distribution centered at 0 with a standard deviation of 0.01. Running these new simulations, we recover the results of the section on No mutations for a variety of parameter settings.

25. See, for example, the large literature on indirect reciprocity. Some of the more famous articles that assume a gossip network in the background are Nowak and Sigmund (1998), Ohtsuki and Iwasa (2004), and Brandt and Sigmund (2004).
26. This is the case even if we limit the initial tolerance levels by drawing instead from a more restricted discrete uniform distribution of 0 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.
27. She will still always adopt the tolerance level of the imitation partner who outperformed him/her.
28. 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) in this journal examine a model of agents with permanent traits ($p = 0$) playing the hawk–dove game. We are not aware of any work that investigates intermediary cases where p is between 0 and 1.
29. When $p = 0$ (that is, traits are completely fixed and cannot be changed) and there are no unconditional hare hunters in the initial population, the population goes neither to universal stag hunting nor to universal hare hunting. Instead individuals have moderate levels of tolerance, making them able to cooperate with those near them in trait space but not able to cooperate with all. Not surprisingly, as the number of unconditional hare hunters initially in the population increases, fewer and fewer of the interactions between individuals in the population consist in both hunting stag.
30. 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).
31. 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 of 0 and a standard deviation of 0.1. We allow tolerance levels to go below 0 and restrict our tags to the interval [0, 1]. Riolo et al. (2001) 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.
32. 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.
33. In these simulations, there was a 10 percent chance for both trait and tolerance mutations, mutations for both traits and tolerance were sampled from $N(0, 0.05)$, and the initial tolerance levels were determined by the uniform distribution from 0 to ½.
34. 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.
35. Peter Singer (1981: 120), *The Expanding Circle*. While we and Singer 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), it is 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.

36. On Friedman's account change occurs slowly as well but for different reasons. Change is slow because it takes time for discriminatory firms to be weeded out of the market. On our account, change is slow because becoming *too* tolerant can be disadvantageous to the individual. At any given time period, there is an upper bound on how tolerant one could profitably be.
37. We are assuming that social contracts differ just in how efficient they are. Of course, some social contracts could be more equitable than others. Whether traits and similarity-based strategies can effectively facilitate transitions from inefficient to efficient social contracts when we take fairness and equity into account will not be explored in this article.
38. For an interesting assessment of this thesis, see Smead and Huttegger (2011).

References

Antal T, Ohtsuki H, Wakeley J, et al. (2009) Evolution of cooperation by phenotypic similarity. *Proceedings of the National Academy of Sciences of the United States of America* 106: 8597–8600.

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

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

Axelrod R and Hammond R (2006) The evolution of ethnocentrism. *The Journal of Conflict Resolution* 50: 926–936.

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: Epstein J (ed) *Generative Social Sciences: Studies in Agent-Based Computational Modeling*. Princeton: Princeton University Press, pp. 177–195.

Binmore K (1994) *Game Theory and the Social Contract, Vol 1: Playing Fair*. Cambridge: MIT Press.

Binmore K (2005) *Natural Justice*. Oxford: Oxford University Press.

Brandt H and Sigmund K (2004) The logic of reprobation: assessment and action rules for indirect reciprocity. *Journal of Theoretical Biology* 231(4): 475–486.

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

Friedman M (1962) *Capitalism and Freedom*. Chicago: The University of Chicago Press.

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

Glaeser E, Laibson D, Scheinkman J, et al. (2000) Measuring trust. *The Quarterly Journal of Economics* 155: 811–846.

Grim P, Selinger E, Braynen W, et al. (2005) Modeling prejudice reduction: spatialized game theory and the contact hypothesis. *Public Affairs Quarterly* 19: 95–125.

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*. In: Curley E (ed.). Indianapolis: Hackett Publishing Co.
Hume D (2000) *A Treatise of Human Nature*. Norton DF and Norton MJ edition (eds). Oxford: Oxford University Press.
James A (2012) *Fairness in Practice: A Social Contract for a Global Economy*. New York: Oxford University Press.
Jansen V and Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440: 663–666.
Kavka G (1986) *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.
Kavka G (1995) Why even morally perfect people would need government. *Social Philosophy and Policy* 12: 1–18.
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.
Muldoon R, Borgida M, and Cuffaro M (2012) The conditions of tolerance. *Politics, Philosophy and Economics* 11: 322–344.
Nowak M (2006) Five rules for the evolution of cooperation. *Science* 314: 1560–1563.
Nowak M and Sigmund K (1998) The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology* 194(4): 561–574.
Ohtsuki H and Iwasa Y (2004) How Should We Define Goodness? – reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* 231(1): 107–120.
Pollack G (1989) Evolutionary stability of reciprocity in a viscous lattice. *Social Networks* 3: 175–212.
Queller D, Ponte E, Bozzaro S, et al. (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.
Robson A (1990) Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology* 144(3): 379–396.
Rousseau J (1984 [1755]) *A Discourse on Inequality*, Cranston M (trans.). London: Penguin Books.
Singer P (1981) *The Expanding Circle: Ethics and Sociobiology*. Princeton: Princeton University Press.
Skyrms B (1996) *The Evolution of the Social Contract*. New York: Cambridge 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: 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.

Smith J and Szathmary E (2000) *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford: Oxford University Press.

Vanderschraaf P (2006) War or peace? A dynamical analysis of anarchy. *Economics and Philosophy* 22: 243–279.

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.

Author biography

Justin P Bruner is a postdoctoral research fellow in the School of Philosophy at the Australian National University. His research focuses on social and political philosophy, social epistemology, experimental philosophy and the mathematical modelling of human and animal behaviour.