PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Forber P, Smead R. 2014 The evolution of fairness through spite. *Proc. R. Soc. B* **281**: 20132439.
http://dx.doi.org/10.1098/rspb.2013.2439

Received: 17 September 2013
Accepted: 15 January 2014

Subject Areas:
evolution, theoretical biology, behaviour

Keywords:
evolution, fairness, spite, assortment, ultimatum game

Author for correspondence:
Patrick Forber
e-mail: patrick.forber@tufts.edu

Royal Society Publishing

The evolution of fairness through spite

Patrick Forber[1] and Rory Smead[2]

[1]Department of Philosophy, Tufts University, Medford, MA 02155, USA
[2]Department of Philosophy and Religion, Northeastern University, Boston, MA 02115, USA

The presence of apparently irrational fair play in the ultimatum game remains a focal point for studies in the evolution of social behaviour. We investigate the role of negative assortment in the evolution of fair play in the ultimatum game. Spite—social behaviour that inflicts harm with no direct benefit to the actor—can evolve when it is disproportionally directed at individuals playing different strategies. The introduction of negative assortment alters the dynamics in a way that increases the chance fairness evolves, but at a cost: spite also evolves. Fairness is usually linked to cooperation and prosocial behaviour, but this study shows that it may have evolutionary links to harmful antisocial behaviour.

1. Introduction

The ultimatum game provides a standard model for studying fair behaviour [1–4]. The game involves two players: a proposer and a responder. The proposer suggests a division of a resource. The responder decides whether to accept the proposal. If the responder accepts, then both players receive the pay-off specified by the proposer, otherwise, neither player receives anything. The game can model fairness, because the proposer has the option to make fair offers, splitting the resource evenly, or unfair offers, biasing the pay-offs in her favour. The rejection of unfair offers by the responder is often interpreted as a punishment inflicted on the proposer for deviating from fairness. The expectation in rational decision settings, and many evolutionary settings, is that individuals should make unfair offers and accept any positive offer in return.

This expectation arises from the game's unique *subgame perfect* solution. Assuming that the responder is rational and seeks to maximize her pay-off, then she should accept any offer that gives her a positive pay-off whether fair or unfair. And, given the rational response, it is rational to demand as much as possible. Experimental results, however, do not cohere with the expectations based on this reasoning, because subjects frequently make fair offers and reject unfair offers [5–7]. Explanations for the experimental behaviour in the ultimatum game often appeal to the evolutionary dynamics of strategies in the game. Despite the game's simplicity and unique 'rational' solution, there are numerous potential evolutionary solutions. Many models show that fairness is a possible evolutionary outcome for this game [8–10]. Given the sacrificial nature of making fair offers and rejecting unfair offers, the evolution of fair behaviour is naturally linked to altruism [11,12]. By contrast, our study shows that the connection is not so straightforward—fairness may have darker evolutionary roots.

Positive assortment facilitates the evolution of altruism; if altruists tend to interact more frequently with other altruists, then such behaviour can avoid subversion by free riders and evolve by natural selection [3,13–15]. Similarly, negative assortment facilitates the evolution of spite, social behaviour that inflicts harm with no direct benefit to the actor and often at some cost; if harm is more frequently inflicted on different types, then spite can evolve [16–18]. Here, we present an evolutionary model of the ultimatum game to study the effects of introducing positive and negative assortment. While positive assortment has little effect, negative assortment tends to aid the evolution of fair behaviour. This occurs by the promotion of spiteful strategies that can, under negative assortment, destabilize the unfair subgame perfect solution. Fairness evolves because it is harder to spite. Introducing mutation [19,20] exacerbates the effect by destabilizing equilibria consisting of entirely fair behaviour, so that the only stable equilibrium involves a mix of spite and fair behaviour. In this way, mechanisms known for generating spiteful behaviour can facilitate the evolution of (partial) fairness.

While the ultimatum game figures prominently in the social and decision sciences, it can also represent strategic interactions in biological systems when there is asymmetric competition over a resource. Suppose an organism arrives at a patch or territory first. The organism can defend most or all of the patch, or it can defend half. A second organism arrives and can challenge the first, at some cost or risk of injury, or accept the part of the patch left. In this way, the ultimatum game is similar to a hawk–dove game [21] except that moves are sequential instead of simultaneous. Parental investment in offspring—in terms of division of investment costs—is another area in biology where subgame perfect reasoning similar to that in the ultimatum game can be applied [22]. Thus, the results we describe here can apply to a large range of strategic settings in biological, social and cultural contexts.

2. The model

To examine the ultimatum game in an evolutionary setting, we consider a single continuous population where individuals are paired randomly to play the ultimatum game. We assume that there is an equal chance of being a proposer or a responder. Given the large number of possible strategies, we focus on a simplified version of the ultimatum game where the proposer may make only one of two offers, fair ($d = 0.5$) or unfair ($1 > d > 0.5$), and the responder has two thresholds for acceptance, any offer or only fair offers (figure 1). Unfair offers (when accepted) mean that the proposer receives d of the resource and the responder receives $1 - d$. Rejected offers result in both parties getting nothing. A strategy in the game specifies what choice to make when in the role of the proposer (fair, unfair) and what choice to make when in the role of the responder (accept, reject). There are four possible strategies in this game: $S1 =$ (unfair, accept any), $S2 =$ (unfair, reject unfair), $S3 =$ (fair, accept any), $S4 =$ (fair, reject unfair). The pay-offs for each strategy pair are given in table 1.

The $S1$ strategy is subgame perfect. The $S3$ and $S4$ strategies both make fair offers as proposer. They differ in their response to unfair offers: $S3$ accepts any positive offer, whereas $S4$ rejects unfair offers. Hence, $S4$ is often described as a fair-minded strategy that punishes deviations from fairness, and $S3$ as a free rider that plays fair, but avoids the cost of punishing. The $S2$ strategy, by making unfair offers as proposer and rejecting unfair offers as responder, displays evolutionary spite. In all interactions, $S2$ acts to minimizes the pay-off of the opponent, often at a cost. For this reason, we call $S2$ the spiteful strategy.

We model evolution of behaviour in this game using the replicator dynamics, where more successful strategies increase in relative frequency [23]. Let $\pi(i, j)$ be the pay-off of strategy i against strategy j, and let x_i represent the frequency of type i in the population. If individuals are paired randomly, then the fitness of strategy i in a population state $X = (x_1, \ldots, x_n)$ is

$$F(i, X) = \sum_j x_j \pi(i, j). \tag{2.1}$$

More successful strategies in the population increase in frequency proportional to the difference in fitness with the population average:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i(F(i, X)) - \sum_j x_j F(j, X)). \tag{2.2}$$

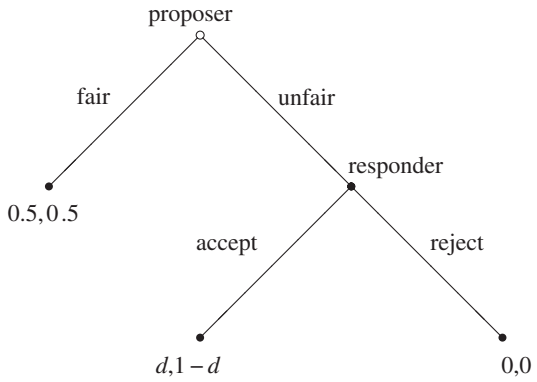

Figure 1. The mini-ultimatum game in extensive form.

Although the subgame perfect equilibrium, a monomorphic population of $S1$, is the only evolutionarily stable state, it is not the only possible outcome of the replicator dynamics. The other possibility is a polymorphic population with some mixture of types $S3$ and $S4$ [8]. These polymorphic states are only stable in a weaker sense; evolutionary drift may carry a population to a point where $S1$ can invade [24,25].

To determine the relative significance of different evolutionary outcomes, we use numerical simulations with a discrete-time version of the replicator dynamics [26] to estimate the basin of attraction for each type of equilibrium. With random initial populations and $d = 0.9$, 71.46% reached the subgame perfect equilibrium. Because subgame perfection, and the concomitant unfair behaviour at equilibrium, is the most likely result, many studies investigate mechanisms that may enhance the probability of reaching states of fair behaviour [4,9,27,28]. We focus on positive and negative assortment between strategies; these factors alter the evolutionary dynamics for the ultimatum game and have unexpected consequences for fair behaviour.

There are a number of mechanisms that can generate positive or negative assortment of strategies: spatial structure [29,30], population structure [31] and conditional strategies based on kinship [13], greenbeards [32] or co-evolving neutral markers [33]. Finite population size also generates some degree of negative assortment [34,35]. Given the range of possible assortment mechanisms, we modelled assortment generically by introducing a uniform exogenous factor that biases the probabilities of interactions between types in the population.

To capture positive assortment, we suppose that an individual interacts with another of the same type with probability r^+, and interacts with a random individual in the population with probability $1 - r^+$. When $r^+ > 0$, there is positive assortment in the population, and the fitness function becomes

$$F(i, X) = (1 - r^+) \sum_j x_j \pi(i, j) + r^+ \pi(i, i). \tag{2.3}$$

For negative assortment, we suppose that individuals interact with a different type with probability r^-, and interact with a random individual in the population with probability $1 - r^-$. When $r^- > 0$, there is negative assortment in the population and the fitness function becomes

$$F(i, X) = (1 - r^-) \sum_j x_j \pi(i, j) + r^- \sum_{j \neq i} \frac{x_j}{1 - x_i} \pi(i, j). \tag{2.4}$$

The factors that bias assortment (r^+ and r^-) involve some important idealizations. For monomorphic populations, negative assortment is not possible. In these cases, we assume that

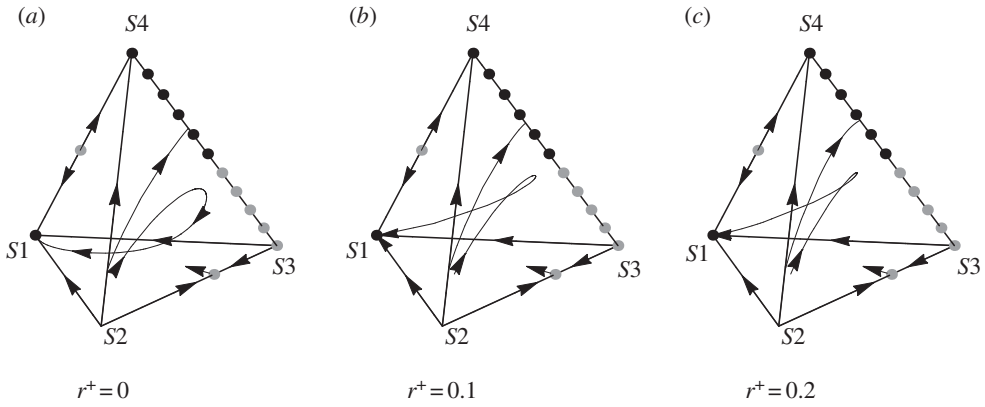

Figure 2. Increasing the degree of positive assortment in the population changes the evolutionary dynamics in the mini-ultimatum game ($d = 0.9$). (*a*) The evolutionary dynamics with random interaction. (*b*) The dynamics with a positive assortment of $r^+ = 0.1$. (*c*) The dynamics with $r^+ = 0.2$.

Table 1. The pay-offs for the mini-ultimatum game. ($1 > d > 0.5$) is the unfair demand and 0.5 out of 1 is the fair demand. Assume all strategies play once in the role of proposer and once in the role of responder.

	S1	*S2*	*S3*	*S4*
S1 (unfair, accept any)	1	$(1 - d)$	$d + 0.5$	0.5
S2 (unfair, reject unfair)	d	0	$d + 0.5$	0.5
S3 (fair, accept any)	$(1 - d) + 0.5$	$(1 - d) + 0.5$	1	1
S4 (fair, reject unfair)	0.5	0.5	1	1

there is no pay-off received from interactions owing to negative assortment. Also note that while the fitness functions are frequency-dependent, the assortment factors are frequency-independent and uniform across types. This idealization permits a tractable analysis of the effects of positive and negative assortment on the evolutionary dynamics. However, it may not be realistic for all mechanisms of assortment. For instance, if assortment is due to individuals actively seeking out others of different types, or if some types negatively assort more effectively than others, then the assortment factor itself (r^-) may vary across types or exhibit frequency dependence. Such factors are important to consider for a complete account of the evolutionary effects of negative assortment, but would significantly complicate the analysis.

3. Results

Positive assortment promotes the evolution of altruism. The best-known instance of this is Hamilton's rule. Suppose some strategy pays a cost c to confer a benefit b on another individual. Then, the condition for the spread of the altruistic strategy is $r^+ > c/b$. This is formally equivalent to Hamilton's rule, because r^+ need not represent any measure of genetic relatedness; it is simply correlation between behavioural strategies [16,31].

Interestingly, positive assortment has little effect on the evolution of behaviour in the mini-ultimatum game (figure 2). There are two types of evolutionary outcomes: a monomorphic population consisting of only *S1* and some polymorphic mixture between *S3* and *S4*. The former corresponds to the subgame perfect strategy, the latter to a set of Nash equilibria where fair behaviour is maintained by threat of punishment. Discrete-time simulations also show that the basins of attraction for the two types of equilibria remain very close in size as positive assortment increases—between 71% and 73% of initial populations reached the subgame perfect state for values of r^+ between 0 and 0.3.

Negative assortment, on the other hand, promotes spite. Suppose that some strategy pays a cost c to inflict a harm h on another individual. The condition for the spread of the spiteful strategy is $r^- > c/h$. The ultimatum game, while a standard for modelling fairness, has a connection to spite: when the responder rejects any positive offer, she pays a cost to inflict a cost. For the responder in the mini-ultimatum game, the cost of spite is the amount rejected $(1 - d)$, and the harm done to the proposer is the demand lost (d). Without negative assortment, *S2* is eliminated by selection, and for this reason, it is often ignored in evolutionary analyses of the ultimatum game. However, *S2* becomes increasingly important as negative assortment is introduced. This opens the door for spite to evolve, and has some unexpected effects on the evolution of fair behaviour in the game.

The subgame perfect strategy is destabilized by spite with sufficient r^-. In addition, for a range of r^- values, an *S2*–*S3* polymorphism is evolutionarily stable (see appendix A). In these polymorphic populations, some individuals make fair offers, some make unfair offers, and there is a significant proportion of rejected unfair offers. Furthermore, if $r^- > (1 - d)/d$, then the subgame perfect strategy *S1* is not stable, and the *S2*–*S3* polymorphism begins attracting the majority of initial populations. The polymorphic mixtures of *S3* and *S4* are also still seen under these conditions, but no other populations are stable. As a result, there are ranges of negative assortment where all evolutionary outcomes involve some degree of fair behaviour even if not all individuals in these populations play fair. For example, when $d = 0.9$, there will be a stable *S2*–*S3* equilibrium whenever $1/14 < r^- < 3/7$, and the *S1* strategy is destabilized when $r^- > 1/9$ (figure 3).

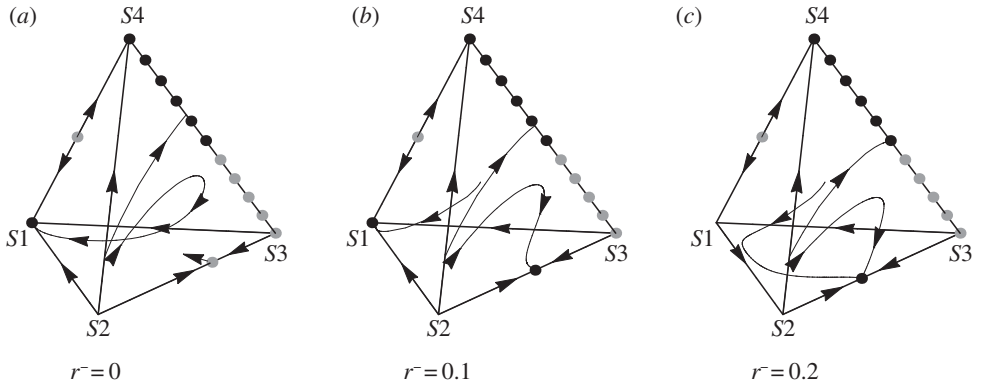

Figure 3. Increasing the degree of negative assortment in the population changes the evolutionary dynamics in the mini-ultimatum game ($d = 0.9$). (*a*) The evolutionary dynamics with random interaction. (*b*) The dynamics with a negative assortment of $r^- = 0.1$. (*c*) The dynamics with $r^- = 0.2$.

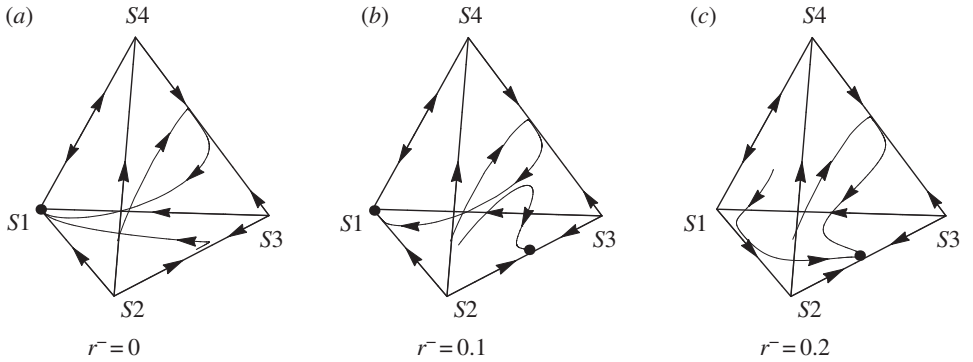

Figure 4. Increasing the degree of negative assortment when including mutation ($d = 0.9$, $\mu = 0.001$). (*a*) The evolutionary dynamics with random interaction. (*b*) The dynamics with a negative assortment of $r^- = 0.1$. (*c*) The dynamics with $r^- = 0.2$.

Table 2. Results from 10 000 numerical simulations on evolving populations. The table shows the frequency of occurrence for each type of equilibrium and the average proportion of fair offers in all simulated populations. $d = 0.9$ and initial populations were chosen randomly and uniformly from the population space.

	$r^- = 0$	$r^- = 0.05$	$r^- = 0.10$	$r^- = 0.15$	$r^- = 0.20$	$r^- = 0.25$	$r^- = 0.30$
all-*S*1 state	0.7146	0.7084	0.5879	0	0	0	0
*S*3–*S*4 state	0.2854	0.2916	0.2732	0.2719	0.2611	0.2605	0.2427
*S*2–*S*3 state	0	0	0.1389	0.7281	0.7389	0.7395	0.7573
% fair offers	0.2854	0.2916	0.3445	0.6060	0.5567	0.5070	0.4374

We used numerical simulations to estimate the basins of attraction for each set of equilibria. As negative assortment increases, the basin of attraction for the *S*3–*S*4 set of equilibria decreases, and the basin of attraction for the *S*2–*S*3 equilibrium increases. However, because both equilibria involve at least partial fair behaviour, there is a nonlinear relationship between r^- and the overall expectation that evolved behaviour will involve fair proposals. In general, evolution in populations with moderate levels of negative assortment leads to an increase in the frequency of fair offers compared with evolution in populations with random interactions. To see this, we can measure the expected frequency of fair behaviour in all evolved populations. Table 2 summarizes these results.

Including mutation in the evolutionary dynamics makes the effect of spite more dramatic. Suppose that, with probability μ, reproduction results in a random type [19,20]. If each strategy is equally likely to result from mutation, then the *S*3–*S*4 populations are no longer stable and tend towards an equal mixture. Additionally, the presence of a small amount of *S*1 and *S*2 causes *S*3 to become more prominent than *S*4 and eventually leads to the evolution of *S*1. This means that, with low values of r^-, populations always evolve to a point close to the subgame perfect strategy *S*1. With higher values of r^-, populations always evolve to a point near the *S*2–*S*3 equilibrium. In such cases, spite is the only thing that can maintain a substantial proportion of fair offers (figure 4).

4. Discussion

Introducing negative assortment of strategies—conditions that favour the evolution of spite—often increases the amount of fair behaviour in evolved populations. Rejecting a positive offer inflicts a harm on the proposer at a cost to the responder: the proposer receives nothing rather than d, and the responder receives nothing rather than $(1 - d)$. Rejection of unfair offers can generate a relative advantage, if

the degree of negative assortment is sufficiently high ($[r^- > (1-d)/d]$). Put another way, negative assortment makes fairness more likely, because spite cannot generate the same relative advantage against fair behaviour. Rejection of fair offers, if they were to occur, could not generate any relative advantage, because the cost would equal the harm. However, negative assortment promotes fairness at the cost of also promoting spiteful behaviour. For most evolutionary outcomes, only partial fairness exists at equilibrium, because many individuals both make and reject unfair demands.

By contrast, negative assortment does not have this effect on a similar division of resource game, the Nash demand game [36]. In the Nash demand game, both players make simultaneous demands. If the demands equal or fall short of the value of the resource, then each player receives their demand; if the demands exceed the value, neither player receives anything. In this game, positive assortment aids the evolution of fair division [3]. The same is not true in the ultimatum game where positive assortment has little effect on the chance of avoiding the subgame perfect equilibrium. The subgame perfect strategy plays well against itself: two *S*1 individuals will exploit one another depending on the proposer–responder roles and, if the probability of assuming each role is equal, do as well on average as two *S*4 individuals playing one another. Instead, negative assortment and the spiteful *S*2 strategy enhances the prospects for the evolution of fairness in the ultimatum game by destabilizing the *S*1 equilibrium. While prosocial behaviour is often linked to positive assortment in other settings, the opposite promotes prosocial behaviours in the ultimatum game, though in a limited way.

Recall that our model represents negative assortment using a uniform exogenous parameter (r^-). This formal representation abstracts away from the underlying mechanisms, though it may capture some more faithfully than others. For instance, small populations generate some negative assortment, because individuals do not interact with themselves. Thus, the distribution of strategies among potential interaction partners differs slightly from the distribution of strategies in the global population by degree $1/(N-1)$ for a population of size N [16]. Negative assortment generated from small populations in this way, or by other external ecological factors, fits well in our model with a constant r^-. However, other mechanisms of negative assortment, such as conditional behaviour or partner choice, can be frequency-dependent or may themselves evolve over time [33,37]. How to extend our model to capture these nuanced mechanisms, and explore the evolutionary dynamics when negative assortment coevolves with spiteful behaviour, are open questions for future research.

The evolution of fairness through spite has the potential to make a significant impact on evolutionary investigations, particularly for human evolution. First, the negative assortment generated by small populations may have had an effect. Given that there is reason to believe that early humans lived in relatively small groups [38], intragroup evolutionary pressures may involve the effects of negative assortment of strategies. Furthermore, humans are cognitively capable of distinguishing in-group and out-group interactions, and can adopt behaviours conditional on group membership. Out-group interactions may involve significant degrees of negative assortment, which may be connected to the in-group favouritism and out-group negativity observed in social psychology [39].

Second, spite has some interesting and complex connections to punishment, a factor of importance in the evolution of social behaviour [40]. In the context of the ultimatum game, spite can be distinguished from punishment. Rejection of unfair offers by the *S*4 strategy is usually interpreted as a punishment that can stabilize fair behaviour in the ultimatum game, though there is some disagreement over whether such punishment counts as altruistic [11] or spiteful [41]. By contrast, such rejection is not readily interpreted as punishment in the polymorphic *S*2–*S*3 populations. It is natural to think of punishment as conditional harming behaviour that is directed at individuals in violation of a behavioural expectation. Punishment often incentivizes cooperative prosocial behaviour. By contrast, spite need not be directed at violations of an expectation, nor incentivize prosocial behaviour.

To see this, consider the two possible polymorphic equilibria in our model. In the stable *S*3–*S*4 populations, fair behaviour is the norm. Rejections are observed only when a strategy is introduced that behaves contrary to this expectation. The *S*4 rejection of unfair offers incentivizes fair behaviour in the ultimatum game in this context. However, in the *S*2–*S*3 populations, there is no such expectation of fair behaviour, because unfair offers are common (as are rejections). Rejections in the *S*2–*S*3 populations are instances of a behaviour that evolved owing to negative assortment. In evolutionary trajectories that lead to the stable *S*2–*S*3 state, rejection involves paying a cost to inflict harm, and thereby reaping a relative advantage. The state is stable, because harming behaviour is disproportionally directed towards any invading types even when no behavioural differences may be observed. Thus, this equilibrium is not maintained by threat of punishment, but by the advantages of spite created through negative assortment.

Of course, both spite and punishment involve inflicting harm, often at a cost to the actor, and so we should expect close evolutionary ties between spite and punishment. They may coincide when, for instance, players keep track of reputation [42] or when reciprocity is involved [43]. This suggests an important role for spite in the explanation of the origin of punishment: spiteful strategies that evolve in single interaction contexts may act as the basis for punishment in more complex contexts. For instance, strong reciprocity explanations for the evolution of cooperation emphasize the stabilizing role of punishment [11]. These explanations, particularly for observed fair behaviour in the ultimatum game, have recently received criticism [44,45]. Spite provides new evolutionary possibilities for explaining the evolution of fairness in these strategic settings.

Finally, because the ultimatum game can apply broadly to asymmetric resource competition in social, cultural or biological settings, we need to countenance the possibility that negative assortment and spite may play important roles in any such interaction. This may even have consequences for the evolution of human morality. Recent approaches often treat prosocial behaviour and its connection to norms of fairness and impartiality as the basis for morality [46]. If morality has its roots in fairness, and spite can facilitate the evolution of fair behaviour in some strategic settings, then antisocial behaviour may have played an important and unsuspected role in the evolution of morality.

Acknowledgements. Both authors contributed equally to this article. Thanks to Kim Sterelny, Brian Skyrms and the audience at the Evolution 2013 meeting for useful comments. P.F. thanks the Syndey Centre for the Foundations of Science at the University of Sydney for support during research for this article.

Appendix A

Considering negative assortment in the population generates a number of interesting new results in the ultimatum game, but also introduces several modelling complications. In what follows, we discuss some of these modelling complications as well as present additional results supporting our claims in the main text.

For negative assortment, we suppose that individuals interact with a different type with probability r^-, and interact with a random individual in the population with probability $1 - r^-$. The fitness of a type i in a population $X = (x_1, \ldots, x_n)$ is given by equation (2.4). The factor $x_j/(1 - x_i)$ normalizes the chance of encountering different strategies after removing i types from the pool. For uniform populations that consist entirely of one type, negative assortment is not possible. In these cases, we assume that there is no pay-off received from interactions owing to negative assortment. A consequence of removing negatively assorted interactions in uniform populations is that fitnesses are only continuous with respect to population space when not at a vertex of the population space. This is one motivation for the introduction of mutation into the evolutionary dynamics. Specifically, the introduction of mutation into the dynamics ensures that all types are present to some degree in any population and avoids potential nuances related to the fitness discontinuities at the vertices.

In the absence of mutation, we can assess the (in)stability of populations by considering the fitness effects of introducing a small number of mutants into the population. Thus, the relevant context for fitness calculations are populations nearby the point under consideration. In the current model, our focus is on the destabilization of the subgame perfect $S1$ strategy and the invasion of the spiteful $S2$ strategy. More precisely, we can define a perturbed population and an accompanying notion of invasiveness as below.

Definition A.1. For a population $X = (x_1, \ldots, x_n)$, let $X^{[i,\varepsilon]}$ be the perturbed population $X^{[i,\varepsilon]} = (x_1 - x_1\varepsilon/(1 - x_i)), \ldots, x_i + \varepsilon, \ldots, x_n - x_n\varepsilon/(1 - x_i))$ where $x_i + \varepsilon < 1$.

Definition A.2. A type i is strongly invasive with respect to a population X if and only if there is some δ such that $F(i,X^{[i,\epsilon]}) > \sum_j F(j,X^{[i,\epsilon]})x_j$ for all $0 < \epsilon < \delta$.

Because our current focus is showing that negative assortment renders the subgame perfect strategy $S1$ unstable, it suffices to show that there is another strategy that is strongly invasive. More generally, if a single type is strongly invasive, it follows that any monotonic selection dynamics will lead to an increase in frequency of the invasive type [47]. The proposition below demonstrates that sufficient negative assortment allows the spiteful $S2$ to invade a monomorphic $S1$ population.

Proposition A.1. *If $r^- > (1 - d)/d$, then $S2$ is strongly invasive with respect to a monomorphic population of $S1$.*

Proof. Suppose that $r^- > (1 - d)/d$ and $X = (1, 0, 0, 0)$. Without loss of generality, consider $X^{[S2,\epsilon]}$ for some ϵ. The fitness functions for each type are obtained by using equation (2.4) with a population of $(1 - \epsilon)$ $S1$ and ϵ $S2$:

$$F(S1, X^{[S2,\epsilon]}) = (1 - r^-)((1 - \epsilon) + \epsilon(1 - d)) + r^-(1 - d) \tag{A 1}$$

and

$$F(S2, X^{[S2,\epsilon]}) = (1 - r^-)(1 - \epsilon)d + r^-d. \tag{A 2}$$

Equations (A 1) and (A 2) represent the fitnesses for each type ($S1$ and $S2$) in a predominantly $S1$ population where ϵ invaders of type $S2$ have been introduced. The inequality $F(S2,X^{[S2,\epsilon]}) > F(S1,X^{[S2,\epsilon]})$ reduces to $r^- > (1 - d)/d$. Because $r^- > (1 - d)/d$ by hypothesis, $F(S2,X^{[S2,\epsilon]}) > F(S1,X^{[S2,\epsilon]})$ for any $\epsilon \in (0, 1)$ and $S2$ is strongly invasive to population X.

In addition to destabilizing the subgame perfect $S1$, the presence of sufficient negative assortment created a new evolutionary equilibrium: a polymorphism between $S2$ and $S3$. The equilibrium point between $S2$ and $S3$ can be found by considering $X = (0,x_2,(1 - x_2),0)$, setting $F(S2, X) = F(S3, X)$, and solving for x_2 (in the absence of mutation). This yields

$$x_2 = \frac{1 - 2d + r - 2dr}{-2 + 2r}. \tag{A 3}$$

Furthermore, for moderate values of d and r ($5/8 < d < 1/4$ $(1 + 2\sqrt{2})$ and $0 \leq r^- \leq 1/3$), this point is strongly evolutionarily stable (strategies $S1$ and $S4$ receive strictly worse pay-offs at this point) whenever:

$$\frac{2 - 2d}{2d + 1} < r^-. \tag{A 4}$$

Note that this point is not on a vertex of the population space and, consequently, not subjected to any potential fitness discontinuities mentioned above. For this reason, standard definitions of evolutionary stability (i.e. evolutionarily stable state) will apply. Larger values for r^- mean that the $S2$–$S3$ equilibrium point gets closer to the all-$S2$ vertex. At this vertex, as with the all-$S1$ vertex, a small number of mutant types can invade—specifically, if a small number of $S4$-types are introduced, they will have strictly greater fitness than the native $S2$ types.

To include mutation in the model, we use the replicator–mutator equation, where μ represents the probability of mutation, and all types are equally likely to occur in mutation.

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i((1 - \mu)F(i, X) - \sum_j x_jF(j, X)) + \frac{\mu}{4}. \tag{A 5}$$

The factor $\mu/4$ represents an equal chance of mutation between each of the four strategies (allowing for self-mutation). Such uniform mutation, in the absence of selection, will push the population towards an equal distribution of all types. It will also mean that the equilibria of the system are slightly perturbed due to the constant influx of other types and are in the interior of the population simplex. In some cases, equilibria that were previously stable will become unstable [19,20]. Asymmetric mutations between different types further complicate the dynamics. For instance, if transitions from $S3$ to $S4$ are sufficiently more likely than other transitions, then a predominantly $S4$ equilibrium can become stable. However, without some rationale for biasing some transitions over others, we treat all transitions as equally probable. Figure 4 shows the dynamics when mutation is present for varying degrees of negative assortment.

Table 3. Results from 10 000 numerical simulations on evolving populations with positive assortment and $\mu = 0$. The table shows the frequency of occurrence for each type of equilibrium and the average proportion of fair offers in all simulated populations. $d = 0.9$ and initial populations were chosen randomly and uniformly from the population space.

	$r^+ = 0$	$r^+ = 0.05$	$r^+ = 0.10$	$r^+ = 0.15$	$r^+ = 0.20$	$r^+ = 0.25$	$r^+ = 0.30$
all-*S*1 state	0.7146	0.7174	0.7223	0.7194	0.7218	0.7228	0.7249
*S*3 – *S*4 state	0.2854	0.2826	0.2777	0.2806	0.2782	0.2772	0.2751

Table 4. Simulation results for negative assortment and $\mu = 0.001$. Note that the populations converge to points nearby the normal equilibria due to the constant presence of mutation. Figures rounded to the nearest 0.0001.

	$r^- = 0$	$r^- = 0.05$	$r^- = 0.10$	$r^- = 0.15$	$r^- = 0.20$	$r^- = 0.25$	$r^- = 0.30$
all-*S*1 state	1	1	0.8859	0	0	0	0
*S*2 – *S*3 state	0	0	0.1141	1	1	1	1
% fair offers	0.0011	0.0012	0.0600	0.4662	0.4097	0.3437	0.2678

We are interested in determining whether negative assortment makes fair behaviour more or less likely. In many cases considered there are multiple evolutionary equilibria. To gauge the evolutionary significance of the equilibria, we can consider the range of initial conditions that will lead to each equilibrium—i.e. estimate the basin of attraction for the equilibria. This was done using numerical simulations with a discrete-time version of the replicator dynamic [26]:

$$x_i^{t+1} = \frac{x_i^t F(i, X)}{\sum_j x_j F(j, X)}, \tag{A 6}$$

where x_i^t is the frequency of type i at time t. The discrete-time dynamic has many of the same stability properties of the replicator dynamic, though differs in the limiting behaviour in some cases. The frequencies of initial population were chosen by drawing a random point in the population space with a uniform distribution and simulations were run until no further change was detected (simulations were written in C). To include mutation in the discrete-time dynamics, we use the difference equation:

$$x_i^{t+1} = \frac{x_i^t (1-\mu) F(i, X)}{\sum_j x_j F(j, X)} + \frac{\mu}{4}. \tag{A 7}$$

The simulation results for negative assortment without mutation are presented in table 2. Tables 3 and 4 summarize simulation results for positive assortment without mutation and for negative assortment with mutation. Positive assortment with mutation always leads to an equilibrium point near the subgame perfect strategy *S*1. These results show that positive assortment has very little impact on the evolutionary outcomes in the mini-ultimatum game, whereas negative assortment can have a dramatic effect. When mutation is present, populations converge to one of two equilibria that are nearby the monomorphic All-*S*1 equilibrium or a stable polymorphic *S*2–*S*3 equilibrium.

References

1. Güth W, Schmittberger R, Schwarze B. 1982 An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Org.* **3**, 367–388. (doi:10.1016/0167-2681(82)90011-7)
2. Thaler RH. 1988 Anomalies and the ultimatum game. *J. Econ. Perspect.* **2**, 185–206. (doi:10.1257/jep.2.2.185)
3. Skyrms B. 1996 *Evolution of the social contract*. Cambridge, UK: Cambridge University Press.
4. Nowak MA, Page KM, Sigmund K. 2000 Fairness versus reason in the ultimatum game. *Science* **289**, 1773–1775. (doi:10.1126/science.289.5485.1773)
5. Güth W, Tietz R. 1990 Ultimatum bargaining behavior: a survey and comparison of experimental results. *J. Econ. Psychol.* **11**, 417–449. (doi:10.1016/0167-4870(90)90021-Z)
6. Oosterbeek H, Sloof R, Van De Kuilen G. 2004 Cultural differences in ultimatum game experiments: evidence from meta-analysis. *Exp. Econ.* **7**, 171–188. (doi:10.1023/B:EXEC.0000026978.14316.74)
7. Henrich J *et al*. 2005 'Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* **28**, 795–815. (doi:10.1017/S0140525X05000142)
8. Gale J, Binmore KG, Samuelson L. 1995 Learning to be imperfect: the ultimatum game. *Games Econ. Behav.* **8**, 56–90. (doi:10.1016/S0899-8256(05)80017-X)
9. Rand DG, Tarnita CE, Ohtsuki H, Nowak MA. 2013 Evolution of fairness in the one-shot anonymous ultimatum game. *Proc. Natl Acad. Sci. USA* **110**, 2581–2586. (doi:10.1073/pnas.1214167110)
10. Fowler JH, Christakis NA. 2013 A random world is a fair world. *Proc. Natl Acad. Sci. USA* **110**, 2440–2441. (doi:10.1073/pnas.1222674110)
11. Gintis H, Bowles S, Boyd R, Fehr E. 2003 Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172. (doi:10.1016/S1090-5138(02)00157-5)
12. Tomasello M, Vaish A. 2013 Origins of human cooperation and morality. *Annu. Rev. Psychol.* **64**, 231–255. (doi:10.1146/annurev-psych-113011-143812)
13. Hamilton WD. 1964 The genetical evolution of social behaviour. I. *J. Theor. Biol.* **7**, 1–16. (doi:10.1016/0022-5193(64)90038-4)
14. Hamilton WD. 1964 The genetical evolution of social behaviour. II. *J. Theor. Biol.* **7**, 17–52. (doi:10.1016/0022-5193(64)90039-6)
15. Fletcher JA, Doebeli M. 2009 A simple and general explanation for the evolution of altruism. *Proc. R. Soc. B* **276**, 13–19. (doi:10.1098/rspb.2008.0829)

16. Hamilton WD. 1970 Selfish and spiteful behavior in an evolutionary model. *Nature* **228**, 1218–1220. (doi:10.1038/2281218a0)
17. West SA, Gardner A. 2010 Altruism, spite, and greenbeards. *Science* **327**, 1341–1344. (doi:10.1126/science.1178332)
18. Smead RS, Forber P. 2013 The evolutionary dynamics of spite in finite populations. *Evolution* **67**, 698–707. (doi:10.1111/j.1558-5646.2012.01831.x)
19. Hadeler KP. 1981 Stable polymorphisms in a selection model with mutation. *SIAM J. Appl. Math.* **41**, 1–7. (doi:10.1137/0141001)
20. Hofbauer J. 1985 The selection–mutation equation. *J. Math. Biol.* **23**, 41–53. (doi:10.1007/BF00276557)
21. Maynard Smith J. 1982 *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.
22. Hammerstein P. 2001 Games and markets: economic behavior in humans and other animals. In *Economics in nature: social dilemmas, mate choices and biological markets* (eds R Noë, JARAM Van Hooff, P Hammersteinrs), pp. 1–20. Cambridge, UK: Cambridge University Press.
23. Taylor P, Jonker L. 1978 Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156. (doi:10.1016/0025-5564(78)90077-9)
24. Bomze IM, Weibull JW. 1995 Does neutral stability imply lyapunov stability? *Games Econ. Behav.* **11**, 173–192. (doi:10.1006/game.1995.1048)
25. Hofbauer J, Sigmund K. 1998 *Evolutionary games and population dynamics*. Cambridge, UK: Cambridge University Press.
26. Weibull JW. 1995 *Evolutionary game theory*. Cambridge, MA: MIT Press.
27. Alexander JM. 2007 *The structural evolution of morality*. Cambridge, UK: Cambridge University Press.
28. Zollman KJS. 2008 Explaining fairness in complex environments. *Polit. Philos. Econ.* **7**, 81–98. (doi:10.1177/1470594X07081299)
29. Pollack GB. 1989 Evolutionary stability on a viscous lattice. *Soc. Networks* **11**, 175–212. (doi:10.1016/0378-8733(89)90002-6)
30. Nowak MA, May RM. 1992 Evolutionary games and spatial chaos. *Nature* **359**, 826–829. (doi:10.1038/359826a0)
31. Frank SA. 1998 *Foundations of social evolution*. Princeton, NJ: Princeton University Press.
32. Gardner A, West SA. 2009 Greenbeards. *Evolution* **64**, 25–38. (doi:10.1111/j.1558-5646.2009.00842.x)
33. Lehmann L, Feldman MW, Rousset F. 2009 On the evolution of harming and recognition in finite panmictic and infinite structured populations. *Evolution* **63**, 2896–2913. (doi:10.1111/j.1558-5646.2009.00778.x)
34. Hamilton WD. 1971 Selection of selfish and altruistic behavior in some extreme models. In *Narrow roads of gene land: the collected papers of WD Hamilton, volume 1: evolution of social behaviour* (*1998*), pp. 198–227. Oxford, UK: Oxford University Press.
35. Grafen A. 1985 A geometric view of relatedness. *Oxford Surv. Evol. Biol.* **2**, 28–90.
36. Nash J. 1950 The bargaining problem. *Econometrica* **18**, 155–162. (doi:10.2307/1907266)
37. Skyrms B. 2003 *The stag hunt and the evolution of social structure*. Cambridge, UK: Cambridge University Press.
38. Dunbar RIM. 1992 Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* **20**, 469–493. (doi:10.1016/0047-2484(92)90081-J)
39. Brown R, Gaertner S (eds). 2001 *Blackwell handbook of social psychology: intergroup processes*. London, UK: Blackwell Publishing.
40. Boyd R, Richerson PJ. 1992 Punishment allows for the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)
41. Marlowe FW, Berbesque JC, Barrett C, Bolyanatz A, Gurven M, Tracer D. 2011 The 'spiteful' origins of human cooperation. *Proc. R. Soc. B* **278**, 2159–2164. (doi:10.1098/rspb.2010.2342)
42. Sigmund K, Hauert C, Nowak MA. 2001 Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10 757–10 762. (doi:10.1073/pnas.161155698)
43. Johnstone RA, Bshary R. 2004 Evolution of spite through indirect reciprocity. *Proc. R. Soc. Lond. B* **271**, 1917–1922. (doi:10.1098/rspb.2003.2581)
44. Hagen EH, Hammerstein P. 2006 Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theor. Popul. Biol.* **69**, 339–348. (doi:10.1016/j.tpb.2005.09.005)
45. Yamagishi T *et al*. 2012 Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proc. Natl Acad. Sci. USA* **109**, 20 364–20 368. (doi:10.1073/pnas.1212126109)
46. Baumard N, André JB, Sperber D. 2013 A mutualistic approach to morality: the evolution of fairness by partner choice. *Behav. Brain Sci.* **36**, 59–122. (doi:10.1017/S0140525X11002202)
47. Sandholm WH. 2010 *Population games and evolutionary dynamics*. Cambridge, MA: MIT Press.