(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Journal of Theoretical Biology 289 (2011) 128–135

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Social opportunities and the evolution of fairness

Jean-Baptiste André [a,*], Nicolas Baumard [b]

[a] Laboratoire Ecologie et Evolution, UMR 7625, CNRS—Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France
[b] Philosophy, Politics and Economics Program, University of Pennsylvania, 313 Cohen Hall, 249 South 36th Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 14 June 2011
Received in revised form
22 July 2011
Accepted 26 July 2011
Available online 5 September 2011

Keywords:
Evolution of cooperation
Partner switching
Outside options
Ultimatum game
Moral psychology

ABSTRACT

We model the evolution of the division of a resource between two individuals, according to a bargaining mechanism akin to the ultimatum game, in which a dominant proposer makes an offer that his partner can only accept or refuse. Individuals are randomly drawn from an infinite population and paired two-by-two. In each pair, a proposer is chosen. The proposer offers a division of resources to his partner. If the offer is accepted it is implemented; otherwise both partners pay a cost and move on to the next social opportunity. When the role that individuals play in each interaction is chosen at random, our analysis shows that each individual receives a fraction corresponding to at least $1/2-c$ of the resource at evolutionary equilibrium, where c represents the cost of postponing the interaction. A quasi-fair division thus evolves as long as c is low. We show that fairness, in this model, is a consequence of the existence of an *outside option* for dominated individuals: namely the possibility of playing on terms more favorable to them in the future if they reject the current interaction. We discuss the interpretation and empirical implications of this result for the case of human behavior.

1. Introduction

Cooperative interactions generate surplus benefits and, beyond understanding qualitatively how natural selection has made their existence among non-kin possible, it is also essential to try and understand *quantitatively* how it has shaped the way these benefits are divided. In particular, although the surplus generated by a cooperative interaction can in principle be distributed in infinitely many ways, human beings systematically express a preference for *fair* divisions. We tend to offer specific quantitative shares to our partners, and expect them to do the same. In symmetric interactions, for instance, we expect equal divisions, and we avoid interacting with people who act unfairly by keeping more. Here, we aim to understand how natural selection shapes the division of benefits in social interactions, and in particular how the preference for divisions of a specific kind, which we call fair, may have evolved.

More precisely, we aim to understand the evolutionary rationale for the simplest version of fairness: the division of a common resource into two equal halves. In this aim, and taking a step further relative to most models on cooperation, we assume the occurrence of a cooperative interaction between two players, generating a surplus of constant size, and we seek to understand how the partners distribute this surplus. To do so, we need to specify a bargaining mechanism for the division. If the two partners have the same bargaining power, it is understandable that fairness evolves, as no one can be forced to accept an unfavorable outcome (Rubinstein, 1982). However, the distribution of resources among humans is typically not the outcome of a mere power struggle. Human beings do favor fair outcomes, even in asymmetric interactions in which one dominant player could in principle take all, or a disproportionate share, of the benefits. This is what we aim to understand.

To this end, we consider a particularly simple and maximally asymmetric negotiation mechanism: the ultimatum game (UG; Güth et al., 1982; Camerer, 2003). In the UG, two individuals share a common resource, but one of them (the so-called *proposer*) benefits from a strategic advantage: he has the power to definitely commit to a certain allocation of the resource with no option to change his mind afterward. The other (called the *responder*) has no option but to either accept the proposed offer or reject it and receive nothing (in which case the proposer also receives nothing). In such a situation, both rationality and natural selection (in simple cases) lead to a maximally unfair outcome: the *proposer* keeps virtually all of the resource. Independently of the amount of time and energy each individual may have contributed to the initial production of the resource, the *proposer* keeps all, because the negotiating power is on his side.

Our aim is to understand how natural selection can lead to fairness in this interaction. In particular, we want to consider the fact that individuals always have *outside options* on a market of social partners. In the ultimatum game *stricto sensu*, each time an

* Corresponding author. Tel.: +33 1 44322341.
E-mail addresses: jeanbaptisteandre@gmail.com (J.-B. André), nbaumard@gmail.com (N. Baumard).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.07.031

individual rejects an offer she definitively compromises an opportunity for social interaction, i.e. individuals have to choose between accepting the very offer they are made, or receive nothing. Yet, in reality, social life is made up of a rich diversity of social opportunities which one can choose to take up, or not (see e.g. Aktipis, 2004). Hence, we believe that it is a mistake to consider any given pairwise interaction as isolated from the relevant outside options. It tends to place an exaggerated importance on purely local power asymmetries.

Recently, inspired in part by the theory of biological markets (Noë et al., 1991; Noë and Hammerstein, 1994, 1995) we have built a model in which proposers make their offer in public and responders can choose the proposer with whom they want to interact before the interaction, thus giving rise to a "social market" (André and Baumard, 2011). Our analysis shows that a fair division of the resource can evolve, provided individuals have the option of choosing (i) their partner, and (ii) the role they wish to play (proposer or responder). One way to understand this result is to realize that the local dominance status of an individual, in a given pairwise interaction, has little influence on the outcome of the interaction if this individual can choose instead to enter into another interaction in which she has a different status.

However, because we initially developed our model to understand the effect of partner choice *per se*, the fact that fairness is fundamentally a consequence of outside options was not easily visible in André and Baumard (2011). Besides, for the sake of simplicity we had to make a number of assumptions. In particular, we assumed that partner choice was perfect and costless. Our aim in the present paper is to develop a further model in which (i) these assumptions are relaxed and (ii) fairness is more clearly shown to be a consequence of outside options in a social market.

To this end, we consider a social interaction based on the ultimatum game. Whereas André and Baumard (2011) considered an idealized paradigm of partner choice, in which responders choose the best among all available offers, in this paper we consider a more parsimonious mechanism based on sequential pairing (see also McNamara et al., 2008). Individuals are randomly drawn from an infinite population and paired two-by-two. In each pair, a proposer is chosen. The proposer offers a resource division to his partner. If the offer is accepted it is implemented. If the offer is rejected, rather than receive a nil payoff, both partners pay a cost (for having postponed their interaction) and move on to the next social opportunity (their "outside option"): they are paired randomly with another partner, and so on.

Importantly, as in Nowak et al. (2000), we assume that the role an individual happens to play in a given interaction is chosen *at random*, i.e. there is no intrinsic property of individuals that is correlated to their probability of being chosen as a proposer/responder. This assumption is meant to represent "social fluidity," i.e. the diversity of social interactions that an individual faces in the course of social life. For the sake of comparison, variations within the same basic model will be considered. In particular, we consider (i) a model in which individuals are stably characterized by a role they play throughout their social life, and (ii) a model in which individuals always remain with the same partner but can change role from one interaction opportunity to the next.

2. General presentation of the model

We consider a simple social interaction based on the ultimatum game (UG). Individuals from an infinitely large population are randomly paired. Each pair of individuals is offered a resource of a given constant size $R=1$ and the opportunity to divide it. One individual in the pair (called the proposer) is strategically dominant, i.e. he is able to propose and commit to a division of the resource, whereas the other (called the responder) has only two options: accept the offer, or reject it (in which case both players receive nothing) and hope for a better social opportunity in the future. Depending on the version of the model, each individual's role is chosen either at random or as a function of intrinsic individual properties. Individuals are genetically characterized by (i) the offer p they make when they play the role of proposer, and (ii) the minimum share of the resource q that they accept when they play the role of responder, called their "acceptance threshold." In a given interaction, the offered split is implemented iff $p \geq q$, otherwise the interaction is canceled (and what occurs next depends on the version of the model).

In the ultimatum game *stricto sensu*, each time an individual rejects an offer she definitively compromises an opportunity for social interaction. This favors undemanding responders, as there is no benefit in rejecting offers. Here, we aim to explore the opposite situation, in which individuals have the option of refusing a social interaction without definitively compromising their chance of interacting later. In other words, we wish to explore the consequences of the existence of a *competition* between various opportunities: rejecting a current opportunity opens up the possibility of accepting another, a form of competition that is absent from the UG. Accordingly, in all versions of the analytical model (but not in individual-based simulations), we assume that the total number of *effective* social interactions each individual undergoes per unit of time is constant (by "effective" we mean an interaction in which the proposer's offer is accepted): it does not depend on the time taken to complete each interaction (e.g. the time it takes to find a compatible partner). Therefore, individuals pay a cost for postponing an interaction in terms of energy consumed and/or in terms of time available for other activities (non-social activities, or other social activities), but not in terms of time available for the very social activity under scrutiny.

In all versions of the analytical model (but not in simulations), we assume that (i) individuals in need of a social interaction enter the population at a constant rate, and that (ii) evolution is slow at the scale of individual lifespan. As a result, the composition of the population of potential partners is considered to be constant across the entire life of an individual.

All analyses assume that mutations are rare, and that recombination is absent (between offer p and acceptance threshold q). As a result, in any evolutionary equilibrium, all the strategies present in the population must reach the same payoff.

The cost of postponing the interaction is measured by a factor $\delta \leq 1$. Consider a social payoff of value 1 obtained immediately. If an offer is rejected and the actual interaction is postponed until the next offer (with the same or a different partner, depending on the version of the model), the very same social payoff will then be worth δ, and δ^2 if the interaction is postponed again, and then δ^3, etc. When $\delta=1$, postponing the interaction is free. When $\delta=0$, postponing the interaction is not an actual option and the game becomes in practice an ultimatum game (refusing an offer is like forgoing any payoff). In practice, our analyses will neglect the situation in which postponing the interaction is completely free ($\delta=1$), as it leads to artifactual neutralities.

Our model assumes that individuals have the option of leaving an interaction without entirely losing the investment they have made in this interaction, once they know (i) which role they play (proposer or responder) and (ii) what fraction of the resource the proposer offers. This is a crucial assumption. It can be interpreted in two different ways. First, the proposer has the ability to commit definitely to a given partition of the resource before the interaction takes place, and the responder then accepts or refuses. Second, the responder has some information on the proposer's usual behavior, either because they actually interact repeatedly (the UG is played several times in a row) or because the proposer

has built up a reputation through past interactions with other individuals. This second interpretation is biologically more realistic. Under this interpretation, we must note that, for the sake of simplicity, our model does not explicitly follow the dynamics of reputation formation. We simply assume that, when an individual is genetically characterized by an offer p, he also has the public "reputation" of offering p.

To confirm analytical results and test their robustness to assumptions, we also developed individual-based simulations in C. The simulation procedure is explained in detail in Section 5 of Supporting Information (SI).

3. The ultimatum game

Before analyzing the model itself, for the sake of comparison, it is worth recalling the evolutionary outcome of the ultimatum game proper (UG), in which individuals have no opportunity to interact further once a given resource division has been rejected.

In the UG, whatever the proposer offers, a responder gains more resources if she accepts it than if she rejects it. Therefore, assuming that natural selection is able to optimize individuals' acceptance threshold (see below), it favors indiscriminate responders, taking whatever resources are made available to them (i.e. $q=0$). Correspondingly, selection favors stingy proposers offering the minimal possible amount, as there is no reason to give more than $\varepsilon \sim 0$ if responders accept it. The evolutionarily stable resource division in the UG is thus be maximally unfair: the proposer keeps all.

Note that, even though the above reasoning is sound, the outcome of evolution in the UG might in fact be slightly less straightforward as there is, in reality, no general reason for responders' acceptance threshold to be optimized by selection (see Gale et al., 1995). This issue is briefly explained in Section 1 of SI. Here we need only note two things.

1. In the UG, there is no evolutionary force pushing the division of the resource toward fairness in particular ($p=0.5$). Depending on the effects of mutations, and on the initial conditions, the evolutionary process can lead to virtually any division of the resource, and there is no particular reason for fairness to be the result.
2. We performed stochastic individual-based simulations with equal mutation rate on offers and acceptance thresholds (see Section 5 of the SI for details on the simulation procedures). For any initial conditions, the average offer and acceptance threshold always evolved toward, and remained very close to 0, in these simulations (Fig. 1).

In the remainder of the paper, these results will be used as a benchmark to which the outcome of the model under different assumptions is compared. We will show that a specific biological mechanism absent from the UG (namely, the option of postponing an interaction in hopes of a better situation in the future) creates a systematic and robust selection force favoring fair or quasi-fair distributions of resources. This will be shown both with analytical arguments and stochastic simulations.

4. Partner switching and social mobility

In the first and primary analysis of this paper, we will consider the general model described in Section 2, in the specific case where (i) individuals systematically leave their current partner when an offer is rejected, and (ii) individuals' role in each social interaction is chosen purely at random. In other words, partner switching and "social mobility" (the option of changing roles) coincide.

In practice, the model works as follows. Pairs of individuals are formed at random. One individual in the pair is *randomly* chosen as the proposer (as in Nowak et al., 2000). He makes an offer p_1 that the responder either accepts or refuses (depending on her acceptance threshold q_2). If the offer is accepted it is implemented, otherwise the two partners are *separated* and move back into the pool of unpaired individuals.

The model aims to consider the evolution of an individual's strategy (p,q), i.e. the joint evolution of individual's offer when playing the role of proposer and acceptance threshold when playing the role of responder. Consider a polymorphic population with a variety of individual strategies, and a focal individual i playing strategy (p_i,q_i).

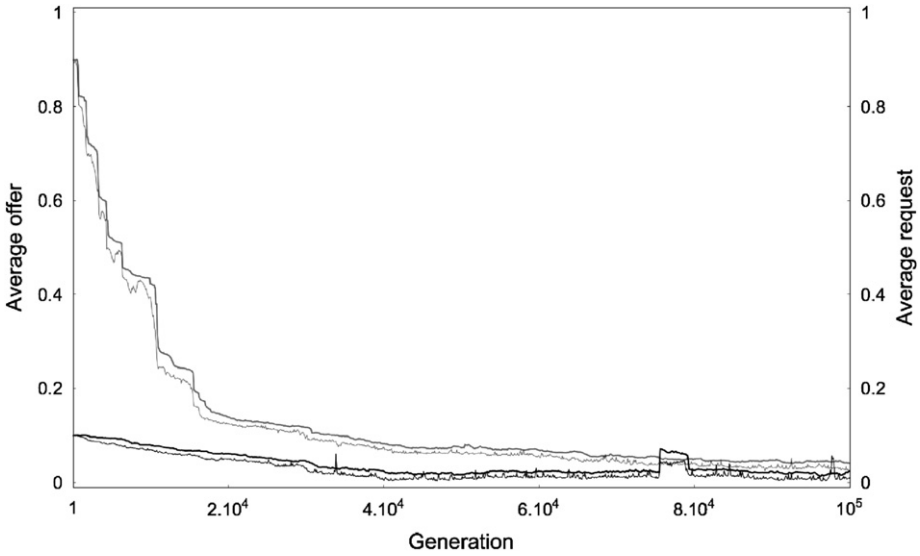

Fig. 1. Ultimatum game. We numerically simulate the evolution of a population of individuals playing the UG (see Section 5 of the SI for details). Each curve is an average over 10 simulation runs. The simulations are initiated in two different ways: (i) with fixed offer and acceptance threshold $p = q = 0.1$ (black lines), (ii) with fixed offer and acceptance threshold $p=q=0.9$ (grey lines). The average offer is indicated by thick lines and the average acceptance threshold by thin lines. The other parameters are the following. The non-social payoff given automatically to each individual is 10^{-5}. The population size is $N=10^3$. The probability of mutation per generation is $\mu=10^{-3}$ on offers and acceptance thresholds. A fraction 0.1 of mutations have a strong effect (the mutated trait is sampled from a uniform distribution between 0 and 1), whereas others have a weak effect (the mutation has a normally distributed effect with standard deviation 0.01).

Call the focal's expected payoff from each effective social interaction G_i. The focal individual is attributed a first partner, and a first role, at random. If this first opportunity is rejected, the focal's effective social interaction is postponed. He ends up exactly in the same situation as before, except that his expected payoff in the next effective social interaction is now only δG_i. As a result, considering every possibility, the expected payoff of the focal individual can be written $G_i = \frac{1}{2}[x(p_i)(1-p_i)+(1-x(p_i))\delta G_i]+\frac{1}{2}[y(q_i)\overline{p}_e(q_i)+(1-y(q_i))\delta G_i]$, where $x(p_i)$ is the fraction of individuals who accept the offer p_i among available partners, $y(q_i)$ is the fraction of individuals who offer at least q_i among available partners, and $\overline{p}_e(q_i)$ is the average offer among them. This gives us the average payoff of individuals playing the strategy (p_i,q_i):

$$G_i = \frac{x(p_i)(1-p_i)+y(q_i)\overline{p}_e(q_i)}{2-\delta(2-x(p_i)-y(q_i))} \quad (1)$$

The explicit evolutionary dynamics of the system are rather complex, in part because the direction of selection on an individual's offer depends upon his acceptance threshold, and vice versa. For instance, if an individual offers $p=0.99$, his expected gain is very low as a proposer, and therefore he should accept almost any offer as a responder, whereas an individual offering $p=0.5$ should be more picky. Such epistasis generates linkage disequilibrium between the two traits (see also McNamara et al., 2008, 2009), rendering the dynamics of adaptation fairly complex to follow analytically.

However, in spite of this complexity, relatively simple analytical arguments can be used to characterize the necessary properties of strategies that can be present in a stable outcome of evolution. In what follows, we describe such a simple analytic argument, and we present the outcome of stochastic individual-based simulations. In the Supporting Information, we develop a more comprehensive mathematical analysis of the system.

4.1. A simple argument

When we described the evolutionary outcome of the UG, we began by following a simple and intuitive argument (Section 3). It consisted in assuming *a priori* that responders' strategy was optimized by natural selection. Because the direction of selection on responders is simple (it always pushes toward a reduction of acceptance thresholds), this led to the conclusion that acceptance threshold was necessarily zero at equilibrium.

As a first approach, we will follow the same simple argument in the present model. We assume *a priori* that responders' acceptance threshold is optimized by natural selection, and derive the necessary properties of a stable end-point of evolution under this hypothesis. The argument unfolds in four steps.

1. *Every individual gains the same G per interaction*: In a population at equilibrium, all individuals must gain the same expected payoff, irrespective of their genotype. Therefore, there exists a payoff G that every individual expects to gain in each social interaction. Note that this payoff is necessarily lower than or equal to 1/2 (individuals cannot expect to receive more than half of the total resource on average).
2. *Every individual accepts exactly δG*: If every individual gains G on average per interaction, irrespective of his *current* partner, one's expected payoff with one's *next* partner is always δG. Accordingly, when playing the role of responder, individuals should always accept offers larger than δG and reject all lower offers. Assuming that natural selection is able to optimize responders' acceptance threshold, the population can be at an equilibrium only if $q=\delta G$ is fixed. Note that this acceptance threshold is strictly lower than 1/2 (because $G \le 1/2$ and $\delta < 1$).
3. *Every individual offers exactly δG*: Individuals should offer exactly the requested amount $q=\delta G$, neither more nor less. This is shown as follows.
   When every individual demands a given $q < 1/2$, individuals can be of three types with regard to their offer. Type 0: some individuals offer $p_0 < q$ and thus obtain an expected payoff $G_0 = \frac{1}{2}\delta G_0 + \frac{1}{2}[y\overline{p}_e+(1-y)\delta G_0] = y\overline{p}_e/(2-\delta(2-y))$, where y represents the proportion of individuals who offer at least q among available partners, and $\overline{p}_e$ is the average offer among them. Type 1: some individuals offer exactly $p_1=q$ and obtain an expected payoff $G_1 = \frac{1}{2}(1-q)+\frac{1}{2}[y\overline{p}_e+(1-y)\delta G_1] = (1-q+y\overline{p}_e)/(2-\delta(1-y))$. Type 2: finally, some individuals offer $p_2 > q$ and obtain $G_2 = (1-p_2+y\overline{p}_e)/(2-\delta(1-y))$.
   First, as long as $p_2 > q$, we have $1-p_2 < 1-q$, and thus $G_1 > G_2$ for any y. Superfluously generous individuals (of type 2) cannot exist at equilibrium. Hence the average offer among individuals who offer at least $q$is exactly $\overline{p}_e = q$ and never more. Replacing $\overline{p}_e$ by q, we can now see that $G_1 > G_0$ for any y, as long as $q < 1-\delta/2$ (which is indeed the case as $q < 1/2$). Therefore, the population can be in a stable state only if individuals of type 1 are fixed, i.e. all individuals offer exactly the requested amount $p=q=\delta G$.
4. *Every individual plays the strategy* $(\delta/2, \delta/2)$: Because every individual in the population offers and requests exactly δG, rejections never actually occur, and the expected payoff of individuals is $G = \frac{1}{2}p+\frac{1}{2}(1-p) = 1/2$. Therefore, at evolutionary equilibrium, the fixed offer is exactly $\hat{p} = \delta G = \delta/2$, and the fixed acceptance threshold is $\hat{q} = \delta/2$.

Note that, like in the UG, even though the above reasoning is sound, the outcome of evolution might in fact be slightly more complex as there is no general reason for responders' acceptance threshold to be optimized by selection. There thus might exist some equilibrium states in which acceptance thresholds are not optimal, because the variability of offers necessary to generate an optimal acceptance threshold is absent. In the Supporting Information (Section 2), we derive the necessary properties of any equilibrium state of the population without the *a priori* assumption that acceptance thresholds are optimized by selection, and we show that our major result essentially holds. In an asymmetric interaction with two roles, when each individual has the option of leaving their current partner and moving on to a new interaction in which they have a chance of playing the opposite role, natural selection leads to an intermediate division, in which every role obtains at least a fraction $\delta/2$ of the resource (i.e. $p \in [\frac{\delta}{2}, 1-\frac{\delta}{2}]$).

4.2. Simulations

Stochastic individual-based simulations were performed (Fig. 2, and see Section 5 of the SI for details on the simulation procedures). They confirm our simple analytical result. With three different values of δ, and two sets of initial conditions, evolution leads to offers and acceptance thresholds that are close to $\delta/2$. Note that when the cost of postponing the interaction is high or low ($\delta=0.3$ and $\delta=0.9$ in our simulations), the average offer is slightly, but significantly, different from what is predicted by our analysis (e.g. $p \approx 0.42$ instead of 0.45), and we see no obvious explanation for this discrepancy.

5. Complementary models

5.1. Partner switching without social mobility

In the main analysis (4), individuals are randomly assigned the role of proposers or responders. There is no such thing as "globally" dominant or subordinate individuals: individuals are

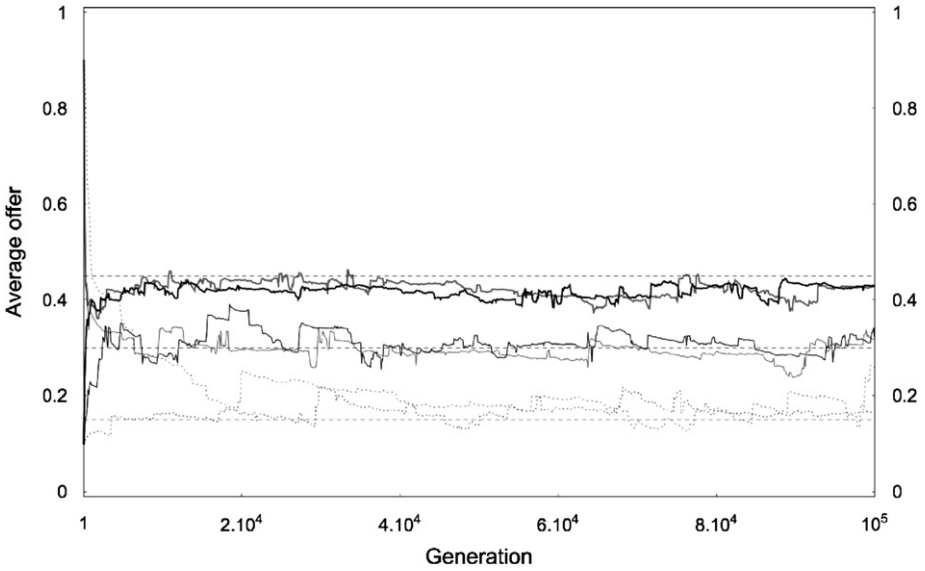

Fig. 2. Partner switching and social mobility. We numerically simulate the evolution of a population of individuals playing the UG when they can change partners and change roles (see Section 5 of the SI). Each curve is an average over 10 simulation runs. The simulations are initiated in two different ways: (i) with fixed offer and acceptance threshold $p = q = 0.1$ (black lines), (ii) with fixed offer and acceptance threshold $p = q = 0.9$ (grey lines). We considered three values of the cost of postponing the interaction: $\delta = 0.9$ (thick plain lines), $\delta = 0.6$ (thin plain lines), and $\delta = 0.3$ (dashed lines). The straight dashed lines show the respective analytical predictions with each value of δ. For every parameter set, we show only the average offer. The average acceptance threshold systematically follows the same pattern (with faster stochastic variation, see Fig. 1 of the SI). The generation length is $L = 10^3$. Individuals who have just interacted socially become "in need" of a social interaction again with a probability $\rho = 0.01$ per time step, and all other parameters are as in Fig. 1.

dominant/subordinate locally (in each given interaction) but not at the "market" level. The market itself is symmetric and gives each individual a fair chance of playing each role. The biological and anthropological relevance of this assumption will be discussed later. For now, for the sake of comparison, we want to explore the consequences of the opposite assumption.

Assume that individuals are assigned a given role at birth (either proposer or responder) that they will always play in every social interaction. For instance, we might imagine that large individuals are always proposers whereas small individuals are always responders. Assume that the frequency of each role is controlled by extrinsic mechanisms (e.g. the occurrence of deleterious mutations) and cannot evolve in response to the payoff obtained in each role (in contrast with André and Baumard, 2011).

To understand the outcome of the model under this assumption, we first follow the same simple argument as in Section 4.1 based on the assumption that responders' acceptance threshold is optimized by selection (see Section 3 of the SI). This argument leads to the conclusion that proposers should always obtain all of the resource at evolutionary equilibrium (i.e. $p = q = 0$). If responders obtain an expected payoff of G_r, they should always request exactly δG_r, i.e. a little bit less than their expected payoff, and this leads to their payoff gradually dwindling to zero.

Note that, here as well, this simple reasoning can be misleading as there is no *a priori* reason that a responder's acceptance threshold must always be optimized by selection. Therefore, here we simply need to note two things:

1. We performed stochastic individual-based simulations (see Section 5 of the SI), which confirm our simple analytical result. With two distinct sets of initial conditions, and an equal mutation rate on offers and acceptance thresholds, we verified that evolution leads to very low offers and very low acceptance thresholds (Fig. 3).
2. In any case, the important point is that there is no selective force favoring fairness or quasi-fairness when individuals cannot change role, i.e. the option of changing partners *per se* does not lead to fairness. If certain individuals play the role *A* in each and every social interaction they can ever undertake, they are forced to accept whatever the average *B* offers them. The emergence of fairness thus crucially relies on the fact that all individuals have a fair chance of playing both roles in each interaction.

Ideally, it would be interesting to consider an intermediate situation, in which certain individuals have a larger probability of being dominant whereas others have a larger probability of being subordinate, but this is beyond the scope of our paper. Here we primarily want to point out the clear-cut difference between two opposite situations. When social role (e.g. dominance) is a global property of individuals, individuals take whatever their average peers also take. When social role is a local property of interactions (but not individuals), individuals can only benefit from a local advantage to the extent that partner switching is costly.

5.2. Social mobility without partner switching

We have shown that partner switching cannot lead to fairness if individuals cannot take advantage of a change in partner to also change roles. But we have not yet tested the effect of the mere option of changing roles without changing partners. In many realistic situations, individuals may have several opportunities to interact, of various kinds, with the same partner, and they may benefit from a local strategic advantage in one interaction but not in another. Is this sufficient to explain the evolution of less imbalanced resource division, or is partner switching *per se* necessary?

To answer this question, we consider the same model as in Section 4 but we assume that, when an offer is rejected by a responder, the two partners remain together and their respective role in the interaction is simply re-attributed at random. Partners are separated only once an actual interaction has taken place (i.e. once an offer has been accepted), except if the interaction turns out to be impossible (because the partners' offers and acceptance thresholds are incompatible), in which case the two individuals are separated (with no payoff). This model is structurally similar to a classic variation on what economists call the infinite-horizon alternating-offers bargaining game (Hoel, 1987; Rubinstein, 1982);

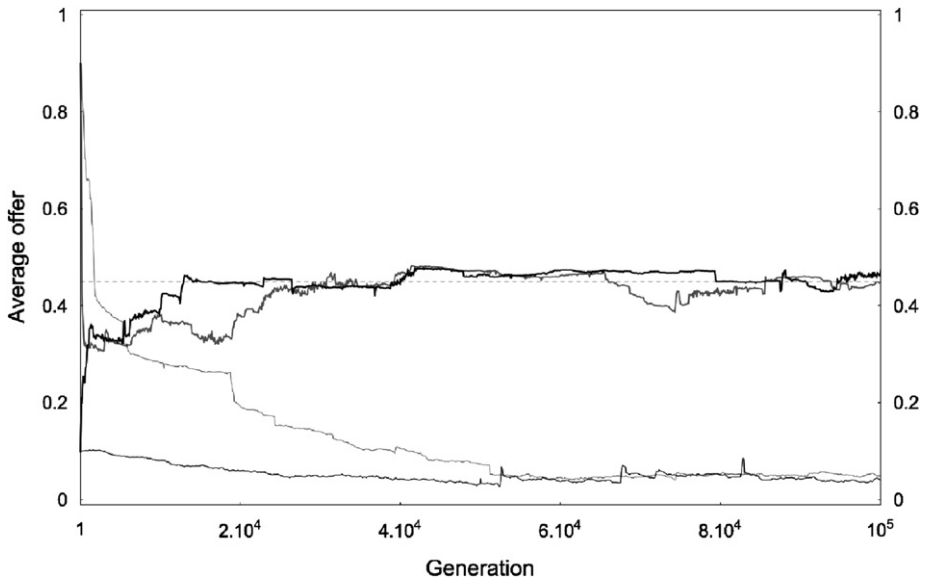

Fig. 3. Complementary models. We numerically simulate the evolution of a population of individuals playing the UG when they can change partners but do not change role (thin lines), and the evolution of a population of individuals playing the UG when they can change role but do not change partners (thick lines); see details in Section 5 of the SI. Each curve is an average over 10 simulation runs. The simulations are initiated (i) with $p = q = 0.1$ (black lines), or (ii) with $p = q = 0.9$ (grey lines). $\delta = 0.9$ in all cases, and all other parameters are as in Fig. 2.

it is known that the only equilibrium strategy in this game that is also resistant to background deviations (i.e. that is "subgame perfect") is a division into two equal halves (when the cost of postponing the interaction is negligible).

Evolutionarily speaking, here we develop a simple argument showing that the only *monomorphic* equilibrium of the population in this game occurs when it is fixed with an offer p and an acceptance threshold $q=p$, with $p \in [\delta/2, 1-\delta/2]$ (see Section 4 of SI). We do not prove that there are no polymorphic equilibria. We also conduct simulations (Fig. 3, and see Section 5 of the SI). They lead to the same results as in Section 4.

The option of changing role, even with a single partner, is thus sufficient to promote the evolution of less imbalanced resource partitions (always within the interval $[\delta/2,\ 1-\delta/2]$). When role is attributed at random at the beginning of each interaction, an individual always expects to gain $\delta/2$ on average in the next interaction (even if it is with the same partner), and therefore should always refuse a payoff of less than $\delta/2$ in the current interaction.

6. Discussion

6.1. Summary of the results

When two partners have the same negotiating power in an interaction, it is understandable that fairness evolves, as no one can be forced to accept an unfavorable outcome (Rubinstein, 1982). However, human beings do favor fair outcomes even in strategically asymmetric interactions in which one dominant player could in principle take all, or a disproportionate share, of the benefits. In this paper, we aim at understanding the evolutionary rationale of these apparently paradoxical preferences.

To do so, we consider a highly asymmetric bargaining interaction, called the ultimatum game (UG), in which one individual (the "proposer") proposes a distribution that the other (the "responder") either accepts or refuses. In this interaction, the proposer benefits from a strong strategic advantage because she is the first to commit definitively to a given offer. We show that subordinate individuals' (responders) having the opportunity to refuse a current interaction, in the hope of playing a different role in a future interaction, is the key lever to overcome this difficulty.

Our model considers a mechanism of sequential pairing. Individuals are sequentially paired with a random partner. In each pair, one individual is chosen to play the role of proposer and makes an offer. If the offer is accepted, it is implemented. If the offer is rejected, the two partners are separated, and paired randomly again. Every individual is genetically characterized by two variables: an offer when playing the role of a proposer and an acceptance threshold (the minimum offer he is willing to accept) when playing the role of a responder. We study the joint evolution of these two variables under the operation of natural selection, both analytically and using individual-based simulations.

In our first, and principal, model, we assume that every individual may play both roles with equal probability (as in Nowak et al., 2000). More precisely, depending on the local situation and/or the partner one is confronted to, one can happen to play either the role of a proposer or the role of a responder, and we assume that the diversity of opportunities is such that, in expectation, each individual has an equal chance of playing each role. In this case, our analysis shows that evolution leads to a division of the resource in which both individuals receive at least $1/2-c$, where c represents the cost paid by both partners when an offer is rejected and the interaction postponed until the next opportunity. Therefore, when the cost of postponing the interaction is very small ($c \simeq 0$), the resource division is quasi-fair (each partner gets almost half of the resource).

To better understand this result, we also develop two complementary models. (i) A model in which individuals are stably characterized by a role that they play throughout their social life, and (ii) a model in which individuals always remain with the same partner but can change role from one interaction opportunity to the next. These models yield a straightforward interpretation. (i) When individuals are characterized by a role that they must play throughout their life, evolution does not lead to fairness, but rather tends to yield highly imbalanced splits in favor of proposers. (ii) When individuals have a chance of playing a different role in the next bout of interaction, even if it is with the same parter, evolution leads to a quasi-fair resource division (each individual obtains at least $1/2-c$). Fairness (or at least less

imbalanced splits) is thus an evolutionary consequence of individuals' having the option of changing roles after they have rejected an offer.

6.2. Interpretation

A simple way to understand these results is to see them as a consequence of *outside options*. It is well known that the existence of outside options constrains the range of equilibrium outcomes in bargaining situations, as no one can be forced to accept less than their outside option (Muthoo, 1999). However, to our knowledge, in game-theoretic analyses, outside options are always fixed a priori, as extrinsic parameters. Here, we consider the fact that, in a market of social opportunities, individuals' outside options are intrinsic to the very interaction under study, because they consist of their payoff in the same interaction, but with different partners, and thus potentially in a different role. In an asymmetric interaction in which both partners have a chance of playing the opposite role in the future, these outside options dramatically constrain the range of equilibrium outcomes. Modulo the cost of playing the outside option (c), the two sides of the interaction must receive the same benefit.

An alternative interpretation of the same results relies on the concept of resource allocation. The marginal value theorem states that, at evolutionary equilibrium, the marginal benefit of a unit of resource allocated to each possible activity (reproduction, foraging, somatic growth, etc.) must be the same (Charnov, 1976). In the social domain this entails that an individual must benefit identically from each resource unit invested into various social endeavors. In particular, when individuals have the opportunity to be on either side in a social interaction involving two distinct roles, then the two must benefit similarly, otherwise one individual is always better off refusing. The human sense of fairness, we think, should thus be understood as a social expression of a general principle that shapes all aspects of resource allocation in living species.

Note that, in the present article, we have considered a social interaction in which each partner invests the *same* amount of time and energy (in fact, we have not even explicitly considered the fact that individuals do invest time and energy, leaving this implicit). In consequence, both partners receive the same benefit at equilibrium. Yet, in further analyses, our results should be generalized to interactions in which partners invest different amounts of resources. In this case, each partner should not be rewarded equally but in proportion to the *opportunity costs* she pays by entering into the interaction. We think that this could help explain the evolutionary origin of an essential feature of morality, whose philosophical formalization dates back to Aristotle: the principle of proportionality.

Chiang (2008) has also developed a model, based on simulations, in which individuals have the option of choosing their partner based on the knowledge of the payoff they have obtained in the past with different partners. He shows that fairness can emerge when the population is initiated with the full range of possible phenotypes (but fairness does not evolve in his simulations when the population is initially composed of "rational" individuals). We believe that the same mechanism must be at work, in a way or another in Chiang (2008)'s model, yielding an equalization of both roles' payoff in the UG. However, it is difficult to arrive at a more precise understanding because Chiang (2008)'s results are only based on simulations.

Our results are also reminiscent of results obtained in the study of reproductive skew in animal societies. Models of skew have long shown that the amount of reproductive opportunities that dominant individuals must leave to subordinates so as to incite them to stay peacefully in the group depends on the subordinates' outside options (Vehrencamp, 1983b; Keller and Reeve, 1994; Nonacs and Hager, 2011). In particular, Vehrencamp (1983a) shows that the imbalance between dominant and subordinate individuals in an animal society is largely reduced if subordinates have a chance of becoming dominant in a future group that they could join. Here, we propose that this very mechanism might explain the evolution of fairness in humans.

6.3. Partner vs. role switching

Fairness emerges in an asymmetric interaction with two roles because individuals have the chance to play the opposite role in the future if they reject the current interaction. From this, we conclude that the possibility of changing roles, not partners, is the key mechanism for the evolution of fairness. But this calls for two comments.

First, on two grounds it may seem to contradict our previous paper, in which we showed (i) that partner choice was necessary for the evolution of fairness, and (ii) that fairness could evolve in a model in which individuals were stably characterized by a given role (André and Baumard, 2011). This is certainly confusing but the contradiction is only a matter of appearances. First, the importance of partner choice in our previous paper was a specific consequence of the assumption that individuals must first choose a role, and then choose their partners. Partner choice is not necessary when individuals simply accept an interaction, in a given role, depending on what they get in it. Second, in our previous paper, we actually showed that fairness required that role frequency be able to change freely in function of the average payoffs of the different roles, which is formally equivalent to the general idea that individuals can preferentially allocate their resources to the more favorable role. In either model, fairness requires a mechanism that relates the effective frequency of a given role to the payoff obtained in this role. The two models are thus more complementary than opposed. However, we believe that the present model offers a more natural way of understanding the principles underlying the evolution of fairness in our day-to-day interactions.

Second, our conclusion that role switching, but not partner switching, is the key mechanism for the evolution of fairness should be understood as a theoretical statement, not an empirical one. In practice, we do believe partner switching to be extremely important. First, it is likely that one's role in relation to a given partner remains the same from one interaction to the next, making partner-switching actually compulsory to change roles. Second, even in scenarios where an individual could in principle play the opposite role later with a given partner, it might still be wiser to change partner in order to increase the chances of a fair outcome. In real-life settings, therefore, it is the opportunity of allocating our time and resources to many different social endeavors, with different partners, that allows us to overcome local power asymmetries. But it is important to understand that fundamentally partner switching has this consequence because it gives dominated individuals a chance of being dominant elsewhere.

6.4. On the cost and option of being choosy

A necessary ingredient for the evolution of fairness in our model is the fact that individuals have the option of rejecting social opportunities, at a moderate cost c, before they are definitely committed to them. At first, this may seem empirically dubious. In natural settings, individuals cannot really make committing offers in advance. Once an "offer" is actually made, it is usually too late to refuse it. Our results thus entail that individuals (here, responders) have some information on the future play of their potential partners before they interact with

them. In the case of human behavior, we believe that this information comes from *reputation*. Others' reputation constitutes information on what one will get if one agrees to interact with them in a given role—i.e., reputation plays the role of a public offer. If someone's reputation does not tell others that they will receive at least the same return on their investment with her as they would receive on average elsewhere, they will refuse to interact with her.

Because reputation conveys information on the likely course of an interaction before one is engaged in it, it can make the cost of "changing" partner very small, and even probably nil in many instances. An individual can mentally screen the reputation of potential partners, and decide with whom to invest time and resources. There are of course exceptions, and one can sometimes be temporarily stuck with an unfair (or incompetent) partner, and have no choice but to accept her offers. However, in real life, the cost of reallocating resources to better social opportunities is likely to vary through time. One may be *temporarily* stuck with an unfair partner, but better opportunities are likely to arise someday. Therefore, it is important to represent the interaction as unfair even if temporarily one cannot do anything about it, in order to continue to actively search for better opportunities (and it is also important for third parties to evaluate the interaction as unfair, to avoid the bad partner).

This suggests an important distinction between behavioral *decisions* and normative *evaluations*. Because it is useful to measure the benefits that we should *aim for*, independently of the benefits we actually end up with, our sense of fairness should not always be in line with our behaviors. This distinction, we contend, could explain the relative universality of normative judgements in spite of the contingency of local situations (see e.g. Marshall et al., 1999).

6.5. Perspectives: unequal outside options

As we have highlighted in several places in the paper, the emergence of fairness in our models depends on the assumption that individuals can change their social role when they change partners. Fairness evolves in spite of local power asymmetries because changing partners creates an opportunity to play a more favorable role, thereby lifting the effect of local asymmetry. We believe that this assumption is realistic in humans, at least in some instances. Human beings participate in an enormous diversity of social interactions, with both a great range of social activities and a great diversity of potential partners involved. They can cooperate in hunting, warfare, cooking, sewing, raising crops, caring for animals, and a great many number of other activities, including entirely novel activities. This rich diversity of situations leaves locally subordinate individuals a chance of finding another social activity in which they are not subordinate.

However, there are certainly exceptions. In real life, all human beings are not always equally totipotent, and there are instances in which an individual's outside opportunities are different from her partner's. Some individuals can be physically stronger than average, thereby being more likely to be dominant. Some individuals may belong to a dominant coalition able to take advantage of others by collectively restricting their opportunities. Some talented individuals may have the ability to produce larger benefits than the average others, thereby having better outside options, which could explain why we intuitively believe that they "merit" to be rewarded more than others. All these issues will require further analyses.

Acknowledgments

We thank Ken Binmore, Nicolas Claidière, Laurent Lehmann, Hugo Mercier, Panayotis Mertikopoulos, Christina Pawlowitsch, Larry Samuelson, and Yannick Viossat for valuable discussions and comments on the manuscript, and we thank two anonymous referees for helpful comments on a previous version of this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2011.07.031.

References

Aktipis, C., 2004. Know when to walk away: contingent movement and the evolution of cooperation. J. Theor. Biol. 231 (2), 249–260.
André, J.-B., Baumard, N., 2011. The evolution of fairness in a biological market. Evolution 65, 1447–1456.
Camerer, C., 2003. Behavioral Game Theory. Experiments in Strategic Interaction. The Roundtable Series in Behavioral Economics. Princeton University Press, Princeton.
Charnov, E.L., 1976. Optimal foraging: the marginal value theorem. Theor. Popul. Biol. 9, 129–136.
Chiang, Y.S., 2008. A path toward fairness : preferential association and the evolution of strategies in the ultimatum game. Ration. Soc. 20, 173–201.
Gale, J., Binmore, K., Samuelson, L., 1995. Learning to be imperfect: the ultimatum game*. Games Econ. Behav. 8 (1), 56–90.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental-analysis of ultimatum bargaining. J. Econ. Behav. Org. 3, 367–388.
Hoel, M., 1987. Bargaining games with a random sequence of who makes the offers. Econ. Lett. 24 (1), 5–9.
Keller, L., Reeve, H., 1994. Partitioning of reproduction in animal societies. Trends Ecol. Evol. 9 (3), 98–102.
Marshall, G., Swift, A., Routh, D., Burgoyne, C., 1999. What is and what ought to be: popular beliefs about distributive justice in thirteen countries. Eur. Sociol. Rev. 15 (4), 349–367.
McNamara, J.M., Barta, Z., Fromhage, L., Houston, A.I., 2008. The coevolution of choosiness and cooperation. Nature 451 (7175), 189–192.
McNamara, J.M., Stephens, P.A., Dall, S.R.X., Houston, A.I., 2009. Evolution of trust and trustworthiness: social awareness favours personality differences. Proc. R. Soc. B: Biol. Sci. 276 (1657), 605–613.
Muthoo, A., 1999. Bargaining Theory with Applications. Cambridge University Press.
Noë, R., Hammerstein, P., 1994. Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. Behav. Ecol. Sociobiol. 35 (1), 1–11.
Noë, R., Hammerstein, P., 1995. Biological markets. Trends Ecol. Evol. 10 (8), 336–339.
Noë, R., Vanschaik, C.P., Vanhooff, J.A.R.A.M., 1991. The market effect—an explanation for pay-off asymmetries among collaborating animals. Ethology 87 (1–2), 97–118.
Nonacs, P., Hager, R., 2011. The past, present and future of reproductive skew theory and experiments. Biol. Rev. 86, 271–298.
Nowak, M.A., Page, K.M., Sigmund, K., 2000. Fairness versus reason in the ultimatum game. Science 289 (5485), 1773–1775.
Rubinstein, A., 1982. Perfect equilibrium in a bargaining model. Econometrica 50 (1), 97–109.
Vehrencamp, S.L., 1983a. A model for the evolution of despotic versus egalitarian societies. Anim. Behav. 31 (AUG), 667–682.
Vehrencamp, S.L., 1983b. Optimal degree of skew in cooperative societies. Am. Zool. 23 (2), 327–335.