Evolution and Human Behavior 37 (2016) 245–254

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Review Article

Models of the evolution of fairness in the ultimatum game: a review and classification

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

[a] Institut de Biologie de l'Ecole normale supérieure (IBENS), INSERM 1024, CNRS 8197, Ecole normale supérieure-PSL Research University, Paris, France
[b] Institut Jean-Nicod (CNRS-EHESS-ENS), Département d'Etudes Cognitives, Ecole normale supérieure-PSL Research University, Paris, France
[c] Institut des Sciences de l'Evolution, UMR 5554-CNRS - University Montpellier 2, Montpellier, France

ARTICLE INFO

Article history:
Initial receipt 13 February 2015
Final revision received 6 January 2016

Keywords:
Human fairness
Ultimatum game
Bargaining problem
Inequity aversion
Equity

ABSTRACT

In the ultimatum game, two people need to agree on the division of a sum of money. People usually divide money equally for the sake of fairness, and prefer to suffer financial losses rather than accept unfair divisions, contradicting the predictions of orthodox game theory. Models aimed at accounting for the evolution of such irrational preferences have put forward a great variety of explanations: biological, cultural, learning-based, human-specific (or not), etc. This diversity reflects the current absence of consensus in the scientific community, and possibly even an absence of debate. Here, we review 36 theoretical models of the evolution of human fairness published in the last 30 years, and identify six families into which they can all be broadly classified. We point out connections between the different families, and instantiate five of the mainstream models in the form of agent-based simulations for purposes of comparison. We identify a variety of theoretical, terminological, and conceptual problems that currently undermine progress in the field. Finally, we suggest directions for future research, and in particular the modeling of the evolution of fairness in a wider and more realistic range of situations.

1. Introduction

In the ultimatum game, two players have to agree on the division of a sum of money. One of the players (called the "proposer") is chosen to make an offer to the other player. The other player (called the "responder") then decides whether or not to accept this offer. If the responder accepts, then both players receive the corresponding sum. But if the responder rejects the offer, neither participant receives any money.

Is it possible to predict what offers humans will make in this game? On the assumptions of orthodox game theory, whereby humans are conceived as well-informed, selfish maximizers, proposers should only make small offers and responders should always accept them. This conclusion simply derives from an application of the rule that "something is better than nothing": since the only alternative to accepting the offer is to receive nothing at all, it is always advantageous for responders to accept low offers. Anticipating that the responder will reason in this way, the proposer should make the smallest possible offer.

However, experimental studies do not confirm this prediction. Güth, Schmittberger, and Schwarze (1982) were the first to test the ultimatum game (UG hereafter) experimentally and to show that offers of 50% are actually very common, while offers below 20% are rejected by responders about half of the time. Researchers have now replicated this seminal experiment hundreds of times, and the original results have held up to all scrutiny. The modal offer in the UG is usually between 40% and 50%, and subjects will reject small offers that are deemed too "unfair" (for a review, see Camerer (2003) or Güth and Kocher (2013) more recently).

How should humans' preference for suffering financial losses rather than accepting unfair divisions of money be explained? Why do humans care more about fairness than about maximizing their monetary payoffs? This behavior is paradoxical not only for traditional game theory, but also for evolutionary biology, which predicts that costly behaviors should not evolve if they do not bring benefits to the individual and/or their genetical relatives in return (Hamilton, 1964; Trivers, 1971; West, Mouden, Gardner, & El Mouden, 2011). Hence, in the last thirty years, a great deal of research has looked for explanations as to why preferences for fairness could have evolved despite their costly effects.

A great diversity of models has been produced. New models continue to be published each year in top-ranked journals, showing that scientific interest in this question is not running out of steam. In fact, if we judge by the journals in which articles are published, the problem of the evolution of fairness is not anymore limited to the fields of evolutionary biology or economics but also tackled by physicists or computer scientists. Despite this profusion of models, it is unclear that our understanding of the origins of fairness is really progressing. Researchers sometimes seem unaware of work related to their own, which suggests a lack of communication. Terminological problems, aggravated by the contribution of scholars from many disciplines,

* Corresponding author. Institut de Biologie de l'Ecole normale supérieure, 46 rue d'Ulm, 75005, Paris, France.
E-mail address: sd@stephanedebove.net (S. Debove).

http://dx.doi.org/10.1016/j.evolhumbehav.2016.01.001
1090-5138/

continue to undermine communication. Theoretical assumptions have become increasingly disconnected from reality. And importantly, no synthesis of the field or cross-model comparisons are currently available.

With these issues in mind, this review has several aims. First, we aim to structure the literature by identifying six families into which all models can be broadly classified. Second, we aim to enhance communication between authors by pointing out the sometimes-hidden connections between models. Third, we aim to identify terminological and theoretical issues that can be easily addressed in order to improve the clarity and consistency of the field. Fourth, we aim to initiate a cross-model comparison, highlighting the weaknesses of models and reproducing five of the major models, coded in the same programming language (we will not analyze those replications here, we only want to make them available to the scientific community at this stage). Finally, we identify promising new directions for future studies.

We may sometimes use the word "fairness" as a shortcut for "fairness in the UG", but our focus is always on the evolution of equal or nearly equal offers in the ultimatum game. Focusing on equal divisions might seem surprising for the reader aware of the wide range of preferences that "fairness" can refer to in everyday life, and we will discuss this problem in Section 4.4. Focusing on the ultimatum game can also seem peculiar as it is only one of many ways to model the division of a resource. In particular, models of bargaining in economics have investigated the division of a resource since at least John Nash's pioneering work in the 1950s (Nash, 1950), long before the term "ultimatum game" was coined. Our focus on the ultimatum game is justified by three points: (1) it is the bargaining game that seems to generate the most cross-disciplinary theoretical work at present, (2) it is a game largely investigated empirically and is thus of interest not only for theorists, (3) the evolution of fairness in the UG has been shown to be more difficult than in related games such as the Nash bargaining game (Alexander, 2007).

Although it is never possible to be entirely exhaustive, we believe that most major models of the evolution of fairness in the UG are present in this review. Models that we deliberately left out of the review are ones where fair preferences are part of the assumptions of the model rather than its outcome (i.e. models that assume non-selfish utility functions such as in Bethwaite and Tompkinson (1996), Fehr and Schmidt (1999), Kirchsteiger (1994)). The most famous model of this kind might be the inequity-aversion model by Fehr and Schmidt (1999)), which shows how a utility function incorporating some preferences for equal outcomes might explain the behaviors observed in the UG. As these models do not deal with the question of how those preferences came to exist in the first place, we do not discuss them. Similarly, we do not review models using axiomatic approaches (assuming pareto-optimality for instance) or studies of the stability of fairness under mutations once fairness has evolved (Harms, 1997; da Silva, Kellermann, & Lamb, 2009, and see SI section 3.1). Readers interested in these modeling approaches and historical models of bargaining more generally can consult the books by Binmore (2005), Skyrms (1996), or Alexander (2007).

Different authors have used different words to name the same strategies in the UG. For example, the minimum offer that a responder is willing to accept can be referred to as a "request", a "demand", an "acceptance threshold", an "aspiration level", or an MAO (for "Minimum Accepted Offer"). Here, we will only use the following terminology: the share of the resource that proposers offer to responders will be referred to as the "offer", and it will be mathematically represented by p. The minimum offer that responders are prepared to accept will be referred to as "acceptance threshold", and will be mathematically represented by q. Some authors also model a simplified version of the UG called the "mini ultimatum game" (mini UG). The only difference between mini UG and classical UG (albeit one that is not necessarily devoid of consequences, as we will discuss in Section 3.2.2) is that the proposer is only allowed to make one of two particular offers: either fair offers of 50%, or selfish offers whose exact value ε varies depending on the authors. Usually, the responder has only two strategies: accept only fair offers, or accept any offer, but some authors have given responders more alternatives (Alexander, 2007).

This review is meant to be mostly non-technical, but a few preliminary considerations may aid in understanding its content. In game theory, it is customary to look for "Nash equilibria". These are sets of strategies for which each player has no interest in choosing a different strategy, knowing the strategies that other players have played. There are an infinity of Nash equilibria in the UG: any situation in which proposers offer responders what they ask for ($p = q$) is a Nash equilibrium (Binmore & Samuelson, 1994). This is because when proposers offer p, responders cannot increase their payoff by increasing or decreasing their acceptance threshold q. At the same time, if responders are ready to stick to an acceptance threshold q, proposers have nothing to gain by making larger or smaller offers. Hence, even the fair strategy ($p = 0.5$, $q = 0.5$) corresponds to a Nash equilibrium. However, this Nash equilibrium can only survive if we assume that lower offers never occur (offers where $p < q$). If a trembling hand or mutations disrupt offers, only the Nash equilibrium where $p = \varepsilon$ and $q = \varepsilon$ (ε close to zero) can survive. This equilibrium is called the "subgame-perfect equilibrium". A related concept that is used more often in biology is the concept of Evolutionary Stable Strategy (ESS) introduced by Maynard Smith and Price (1973). An ESS corresponds to a strategy that cannot be invaded by any vanishingly rare mutant strategy if adopted by all other members of a population. Hence, in the following sections, references to the evolution of Nash equilibria that are usually not subgame perfect, or to ESSs that depart from the selfish one, will be two ways of rephrasing the problem of the evolution of fairness.

2. Six families of models of the evolution of fairness

We classify models according to the mechanism that the authors suggest as the driver of the evolution of fairness, which might be the most obvious criterion. However, some mechanisms identified as different are so similar that it is questionable whether it makes sense to distinguish them. We identify those hidden connections between models in SI section 3.2. For each family of models, we only present what we think to be the most important or seminal papers and explicitly describe the mechanism that allows fairness to evolve (when it is possible to identify it). Table 1 summarizes the classification. SI section 2 presents the classification of more recent models that we could not include here for reasons of space.

2.1. Alternating role-based models

We start with the first, historical models on the evolution of fair offers: alternating-offers models of bargaining (Stahl, 1977; Rubinstein, 1982; Hoel, 1987). Rubinstein (1982) studies the following problem: "Two players have to reach an agreement on the partition of a pie of size 1. Each has to make **in turn** a proposal as to how it should be divided. After one player has made an offer, the other must decide either to accept it, or to reject it and continue the bargaining" (our emphasis). Note that acceptance of an offer ends the bargaining, so this game is different from a repeated UG strictly speaking. Additionally, each player's payoff is multiplied by δ ($0 < \delta < 1$) when entering a new bargaining period (i.e., payoffs are discounted by δ at each new period). Rubinstein (1982) shows that when δ is the same for both players and tends toward 1 (there is no discounting, so rejecting an offer is not costly), the perfect equilibrium of the game is an equal division.

The intuition behind this result is straightforward: there is no reason to accept offers smaller than 0.5 when responders know they will play the role of proposer in the next period. Ultimately, as both players can use this reasoning, the only offer that can be accepted is 0.5.

Hoel (1987) relaxes the assumption of a strictly alternating sequence of offers by assuming that in each round, a random draw determines who gets to make the offer. He shows that fair offers

nevertheless evolve under this less strict mechanism. In fact, the introduction of random roles allows the fair equilibrium to be reached in five periods, in contrast to the game of Rubinstein (1982) which has an infinite horizon.

Hence, having the chance to hold a dominant position in a bargaining interaction some of the time, even randomly, is enough for fair divisions to evolve. Hoel (1987) cites institutional factors such as bureaucratic delays or tactical considerations as the real-life equivalent of this mechanism. Another interpretation could be that human beings, used to varied and repeated interactions in their daily life, have been culturally trained to make fair offers (Gale, Binmore, & Samuelson, 1995; Skyrms, 1996), or have an evolved, biological sense of fairness that they bring and use in the lab.

2.2. Reputation-based models

Nowak, Page, and Sigmund (2000) suggest that reputation may contribute to the evolution of fairness. In their model, individuals play UGs repeatedly, and each time two individuals reach an agreement, a fraction of the population learns about the offer that has been accepted. In subsequent interactions, they will be able to offer whichever is smaller, their own p-value (the offer they are genetically characterized by) or the minimum offer that they know their partner has accepted in the past. Nowak et al. (2000) show that this mechanism is enough to lead to the evolution of fairness, as long as the fraction of individuals who learn about the outcome of any interaction is large enough (Nowak et al., 2000, Fig. 2). However, this result is only possible because the authors make an assumption that drastically restricts the parameter space, as they themselves recognize (Nowak et al., 2000, footnote 14). We will return to the problem of assumptions in Section 3.2.1.

Chiang (2008) presents a reputation-based model in which individuals exhibit partner preferences: they prefer partners who have brought them greater accumulated benefits in the past. Chiang (2008) shows that this type of preferential association can lead to fairness, but is heavily dependent on the initial state of the population. In particular, with an initial population of selfish agents ($p = 0, q = 0$), fairness will not evolve.

André and Baumard (2011) suggest another way that reputation could lead to the evolution of fairness. Whereas in Nowak et al. (2000) proposers use reputation to make smaller offers to responders, in André and Baumard (2011) both proposers and responders use reputation to decide who they should interact with. In other words, reputation serves as a way to *choose* partners, whereas in Nowak et al. (2000) it serves as a way to *control* partners. The need to avoid being left out of interactions prevents proposers from being entirely selfish and requires them to increase their offers, leading to the evolution of fairness. Importantly, a partner choice framework cannot lead offers to increase above 50%. As soon as offers start surpassing 50%, there is less incentive to play the role of proposer than to play the role of responder. Hence, individuals will stop taking on the role of proposers, which will drive responders not to have an acceptance threshold of more than 50% in order to continue finding opportunities to interact. As a result, the only offers and acceptance thresholds at the evolutionary equilibrium are $p = 0.5$ and $q = 0.5$. In partner choice-based modeling, individuals are rewarded according to their outside options: they always end up obtaining the best that they could obtain somewhere else in the population (see also (Debove, Baumard, and André (2015)).

2.3. Noise-based models

Binmore and Samuelson (1994) and Gale et al. (1995) were the first to suggest that noise could explain results from the UG. They consider UG players as agents who can be in one of two modes: playing mode or learning mode. In playing mode, agents choose their strategy for the next UG according to a specific decision rule; in learning mode, agents adjust this decision rule. Gale et al. (1995) show that if the learning mode is noisier for responders than for proposers (for example, because more responders than proposers mistakenly learn a strategy), non-subgame-perfect Nash equilibra can be expected.

The intuition behind this result is straightforward: refusing low offers is costly for responders only if a large number of proposers make low offers. But when responders' behavior becomes noisy enough compared to proposers, it becomes costly for proposers to make low offers, as they have a high probability of being rejected. Soon enough, pressure on responders' to accept low offers becomes negligible compared to noise-induced drift, and proposers have to adapt by increasing their offers.

Roth and Erev (1993) reach a conclusion similar to that of Gale et al. (1995), with a model in which the propensity q to make a particular offer k at time t is determined by the payoff x received with this offer in the previous time period: $q_k(t+1) = q_k(t) + x$. This updating dynamic leads to offers that closely approximate experimental data. The authors interpret this pattern as being driven by an asymmetry of payoffs between responders and proposers. On the one hand, the difference between what proposers gain when their selfish offer is accepted or rejected is large. On the other hand, the difference between what

Table 1
17 models of the evolution of fairness and their main characteristics.

	Mechanism	Timescale of evolution	Human-specificity	Restricted to UG
Alexander (2007)	Spatial population structure	Cultural	N/S	No
André and Baumard (2011)	Reputation and partner choice	Biological	Diversity of social interactions	No
Barclay and Stoller (2014)	Spite (local competition)	Biological	N/S	No
Chiang (2008)	Reputation (Preferential association)	Cultural	N/S	No
Forber and Smead (2014)	Spite	Social, cultural, or biological	N/S	No
Gale et al. (1995)	Acceptance thresholds noisier than offers	Interactive learning	N/S	Yes (learning)
Hoel (1987)	Alternating offers	Economics dynamics	N/S	No
Huck and Oechssler (1999)	Spite	Biological or Cultural	N/S	No
Iranzo et al. (2011)	Spatial population structure	Biological or cultural	N/S	No
Killingback and Studer (2001)	Spatial population structure	Biological	N/S	No
Nowak et al. (2000)	Reputation	Biological or cultural	N/S	Probably No
Page et al. (2000)	Spatial population structure	Biological	N/S	No
Page and Nowak (2002)	"Empathy", $p = q$ assumption	N/S	N/S	N/S
Rand et al. (2013)	Noise (weak selection/high mutation)	Biological or cultural	N/S	Probably (interindividual variation)
Roth and Erev (1993)	Stronger selection on proposers than responders	Learning	N/S	Yes (learning)
Rubinstein (1982)	Alternating offers	Probably short timescale	N/S	No
Zollman (2008)	"Noise" (complex environments)	Probably cultural	N/S	No

Classification can be subjective as authors sometimes do not make their interpretation explicit, but SI Table 1 provides the elements on which our classification is based. N/S = Not Specified. Human-specificity: the authors' explanation for why fairness might be restricted to humans (see Section 4.3). Postulating cultural evolution is not enough if the focus mechanism has no obvious reason to be limited to cultural evolution. Restricted to UG = "No" means that authors think their model can explain fairness outside the UG (see Section 4.2).

responders gain when they accept or reject a low offer is small. As a result, proposers learn that they should not make selfish offers faster than responders learn that they should accept them. In biological terms, selection is stronger on proposers than responders. The mechanism is thus self-reinforcing: once proposers have learned that they should not make low offers, responders have no incentive to learn not to reject them.

In the same vein, a recent model by Rand, Tarnita, Ohtsuki, and Nowak (2013) suggests that weak selection and a high mutation rate can explain the evolution of fairness in the UG. Although their interpretation is not framed in terms of noise, weak selection and a high mutation rate have this effect: they keep reintroducing a variety of different, and sometimes maladaptive strategies, into the population. If demanding responders keep being reintroduced, then proposers can no longer afford to make low offers, and are under a selective pressure to increase their offers.

Finally, Zollman (2008) shows that when agents have to play not only a UG but also a Nash demand game, it helps fairness to evolve. Whether or not this kind of "complex environment" can be said to be noisy is debatable, but we still include this model because it is a good example of trying to model the evolution of fairness in more diverse environments (see Section 4.4 on this point).

2.4. Spite-based models

Huck and Oechssler (1999) suggest that if responders can inflict more costs on proposers than on themselves by rejecting small offers, fairness will be able to evolve. Although they do not use the word, this resembles the definition of "spite" in evolutionary biology (Lehmann, Bargum, & Reuter, 2006; West & Gardner, 2010). It is well known that spite is more effective in small populations, because the relative gain of inflicting costs on others is higher in this situation. Indeed, Huck and Oechssler (1999) find that population size matters: the larger proposers' offers (and thus the higher the cost of refusing them), the smaller the population must be for fairness to evolve.

Forber and Smead (2014) show that introducing negative assortments between the four possible strategies in the mini UG will destabilize the subgame-perfect equilibrium. They find that a mixture of strategies involving fair offers can stabilize, which sometimes include [make unfair offers, reject unfair offers] strategies. They call these strategies "spiteful" strategies. Their interpretation of the evolution of fairness is as above in terms of asymmetry of costs inflicted and costs received: spiteful strategies inflict a larger cost on unfair proposers than on fair proposers.

Barclay and Stoller (2014) also insist on the importance of spite, in a model showing that it pays off to accept offers whenever they are higher than

$$\frac{2}{2N + ak(N-2)} \quad (1)$$

with N being the number of group members, k the size of the resource to be divided, and a the proportion of offers accepted in the population. Hence, as group size increases, or the more people accept offers in the population, the more it pays off to accept small offers. Barclay and Stoller (2014) complement their model with a behavioral experiment showing that, following the predictions of spite-based models, people tend to accept lower offers when they are competing for money with a larger group.

Note that strictly speaking, spite requires special conditions to work, such as negative relatedness between the actor and the recipient. These conditions are thought to be rarely met in nature (West & Gardner, 2010), and supposedly spiteful behaviors can usually be re-described as selfish, in the sense that the short-term cost paid by the actor increases her fitness in the end. As models usually do not detail relatedness, we are unable to know whether they investigate real evolutionary spite or not.

2.5. Spatial population structure-based models

Most of the models cited above assume "well-mixed" populations, in which individuals are randomly drawn from the whole population to play UGs. Page, Nowak, and Sigmund (2000) relax this assumption to study the effects of spatial population structure on the results of the UG. They analyze a model in which agents are arranged either on a ring or a square grid, so that they play UGs and compete for offspring only with a few individuals in the population (their neighbors). They are interested in the conditions that will prevent a mutant from invading the resident population with such a spatial structure. Making the assumption that $p_{mutant} \geq q_{mutant} \geq p_{resident} \geq q_{resident}$, they show that the smaller the neighborhood size, the larger the offers will be at the evolutionary equilibrium.

Killingback and Studer (2001) investigate the same mechanism without the assumption that $p_{mutant} \geq q_{mutant} \geq p_{resident} \geq q_{resident}$, but with the additional assumptions that some agents are more dominant than others and that more dominant agents always play the proposer role. They show through simulations that spatial structure can lead to offers up to 0.45, but their analytical argument is not detailed enough (p. 1800 paragraph 2) to really understand the mechanism at play.

Alexander (2007) studies the evolution of fairness on lattices, small-world networks, bounded-degree networks and dynamic networks, under a variety of initial conditions, mutation rates, etc. It is not possible to summarize this wealth of models in just a few lines, but the general result is that spatial structures do not really facilitate the evolution of fairness. The easiest evolution happens on dynamic social networks, where players update their probabilities of interaction with their neighbors at the end of each generation, but even then fair offers dominate the population only a third of the time and in very special conditions (Alexander, 2007, p. 235). Note that Alexander (2007) uses a mini UG contrarily to the models presented above, which could explain the discrepancy in the results.

Iranzo, Román, and Sánchez (2011) study a spatial UG under a variety of strategy copy mechanisms (in some cases always biased toward the highest payoff, in others not), forms of proposer/responder role assignment (random or alternating), and fidelity of replication (presence or absence of noise). They find through simulations that fair offers can evolve under multiple combinations of such parameters, but their analytical argument assumes that $p < 1/2$ (p. 8, paragraph 2), which makes it impossible to determine whether fair offers are intrinsically advantageous.

It is not exactly clear whether a single mechanism is at play in all spatial UG models. In Page et al. (2000), fairness seems to evolve (our interpretation) because individuals who refuse small offers (1) also cause their direct and only competitor for offspring to miss an opportunity to interact and (2) control their neighbor's payoff through the offer that they make in the only interaction accepted between the pair. Hence, to avoid being at their neighbor's "mercy", individuals have to increase their offers in order to continue playing the role of proposer. This result may rely strongly on the assumption that $p_{mutant} \geq q_{mutant} \geq p_{resident} \geq q_{resident}$, something that the authors do not discuss.

2.6. Empathy-based models

A few models investigate the importance of "empathy" for the evolution of fairness. Empathy means that individuals only make offers that they would themselves be ready to accept (mathematically, $p = q$), or have acceptance thresholds that are not higher than what they would offer themselves ($q = p$). Page and Nowak (2001, 2002) show that allowing a small proportion α of the population to play the empathetic strategy is enough to lead to the evolution of fairness. The authors interpret this result as the result of a selection "pressure for q to increase

in order to avoid rejection" (p. 1110, last paragraph), but it is unclear why the assumption $p = q$ should create more fear of rejection than in the traditional UG without empathy. Additionally, the authors show that if natural selection can act upon α, it will be driven to zero. Hence, "empathy" understood as $p = q$ is not itself selected and must be explained by another mechanism.

Sánchez and Cuesta (2005) investigate the effect of the assumption $p = q$ on the evolution of fair offers, but they also add a large amount of "noise" in their model. It is thus possible that the evolution of fairness they obtain is the result of noise rather than empathy, especially as the distribution of offers that they obtain never reaches a stationary value.

It is thus difficult so far to pinpoint why the assumption $p = q$ leads to fairness; one explanation we might suggest is that it adds noise to the model.

3. Terminological and theoretical problems

3.1. Terminological problems

3.1.1. Loose usage of terms

The first problem is not specific to the field of the evolution of fairness, but to the field of the evolution of cooperation as a whole: terms are used in a loose sense, when they are not simply used in a wrong sense (West, Griffin, & Gardner, 2007; West, Mouden, et al., 2011). This problem is exacerbated by the participation of scholars from many disciplines in the field. For instance, Sánchez and Cuesta (2005) study the evolution of fairness in a regular UG, but switch between talking about "altruism", "strong reciprocity", "altruistic punishment", "other-regarding behavior", and "empathy" in discussing their results. These terms either are not well-defined in evolutionary biology or refer to very different biological realities, so treating them as interchangeable in the same paper can only create confusion regarding what biological trait is being investigated.

Here we can only refer to two excellent and human-oriented reviews by West, Mouden, et al. (2011) and West, Griffin, et al. (2007) for a semantic clarification, and encourage authors to define what they mean by "fairness" if they depart from the traditional definition of (almost) equal divisions found in the UG.

3.1.2. Different definitions of fairness

There is currently no agreed definition of fairness in the social sciences, and the definition in the context of the UG is no clearer, given the wide variability of behaviors observed in the game. The authors of most of the papers that we examined rely on the evolution of offers of 0.4–0.5 to conclude that fairness has evolved. This also corresponds to the modal offer in the empirical UG. Nonetheless, some authors consider fairness to have evolved at much smaller values. Wang, Chen, and Wang (2014) report fairness for offers of 0.35, while Ichinose and Sayama (2014) describe offers as low as 0.25 as fair (see their Fig. 1, p. 2, paragraph 4). Although there is a significant quantitative difference between 0.25 and 0.5, this difference is obscured if papers with such widely differing findings report the evolution of "fairness" in their results or title.

3.2. Theoretical concerns

3.2.1. Putting constraints on offers and acceptance thresholds

Some authors place constraints on offers and acceptance thresholds (p and q) to "help" fairness to evolve. We mentioned in Section 2.2 that the model by Nowak et al. (2000) suffers from one such limitation: the authors assume that the resource left to individuals when their offer has been accepted must not be smaller than what they would ask when playing the role of responder. In other words, the authors restrict the parameter space so that $1 - p \geq q$.

To illustrate the heavy impact of this restriction, we reproduced the model of Nowak et al. (2000) with and without the restriction that $1 - p \geq q$ (see Methods in SI section 4.1). The results are presented in Fig. 1. With $1 - p \geq q$, we replicate the results of Nowak et al. (Fig. 1, circle markers), but without this restriction offers evolve toward the maximum possible level (Fig. 1, triangle markers). This result is easy to understand: when proposers have information on the offers previously accepted by responders, the roles in the UG are actually reversed. Through their reputation, responders are actually the ones to first suggest a division of the resource, and proposers are the ones left in the situation of deciding whether to accept it or to receive nothing at all.

Nowak et al. (2000) are not the only ones to place constraints on the range of offers and acceptance thresholds that can evolve: all empathy-based models assume $p = q$, and Chiang (2007) assumes $p + q = 1$. It is of course part of any modeling process to make simplifying assumptions, but the problem here is that these assumptions arbitrarily restrict the values that can be taken by the very variables whose distribution is to be explained. It is also difficult to justify these restrictions on the basis of their supposed "reasonableness" (Nowak et al., 2000 p. 1773, paragraph 3), as the evolution of "unreasonable" preferences is precisely the subject of any model of the evolution of fairness. Finally, the biological basis of this assumption is unclear: why would it be the case that offers and acceptance thresholds cannot evolve independently? Hence, we suggest that authors have at least one condition in which they allow offers and acceptance thresholds to evolve independently, and take great care when interpreting results obtained by restricting the $[p,q]$ parameter space.

3.2.2. Using a mini-ultimatum game

Using a mini UG is another way to put constraints on p and q. In this case, each variable is only allowed to take one of two values: selfish or fair. This assumption presents at least three problems. First, the numerical value of the "selfish" option differs from one author to another, and no indication is usually given as to how the evolution of fairness depends on this value. Second, the fair-or-nothing nature of the mini UG makes it difficult to interpret biologically. Finally, and most importantly, it is impossible to know whether there is something intrinsically advantageous about making fair offers of 50% or if offers of 30% or 70% would also have outcompeted the unfair offers used in a given study.

Hence, while using a mini UG can be helpful in understanding how a specific mechanism leads to fairness, we suggest using a continuous UG when possible, or at least a UG in which the range of possible offers is discretized with enough values between 0 and 1 (Binmore & Samuelson, 1994; Gale et al., 1995; Harms, 1997), and see Skyrms (1996) for a discussion of discretizing evolutionary games to different degrees).

3.2.3. Empathy modeling

Finding the evolution of the relationship $p \approx q$ in a model is not surprising, as it is usually the case that increasing offers are driven by increasing acceptance thresholds (and natural selection favors proposers who offer just a little bit more than the responders' acceptance threshold). In fact, in almost all the models we reviewed, it is the case that $p \approx q$. This does not constitute a result in itself, and accounts of the evolution of fairness in these models couched in terms of "empathy" at best offer a re-description of the system. Iranzo, Flora, Moreno, and Sánchez (2012) go as far as to declare that their model "could explain the emergence of empathy in very many different contexts" (p. 1) after obtaining the evolution of $p \approx q$, which also requires the acceptance of a restricted definition of empathy as offering to others what one would require for oneself.

If obtaining the relationship $p = q$ does not mean one has explained the emergence of empathy, the reverse is also true: while the emergence of fair offers is helped by $p = q$, can it be concluded that fairness results from empathy, biologically speaking? We understand that it may have been convenient for Page and Nowak (2002) to use the word "empathy" as a shortcut for $p = q$ in explaining their model, but the drawback is that this paper is continually cited as a demonstration that "empathy explains fairness", with no supplementary word of caution. At the very least, we should be very cautious with the affirmation

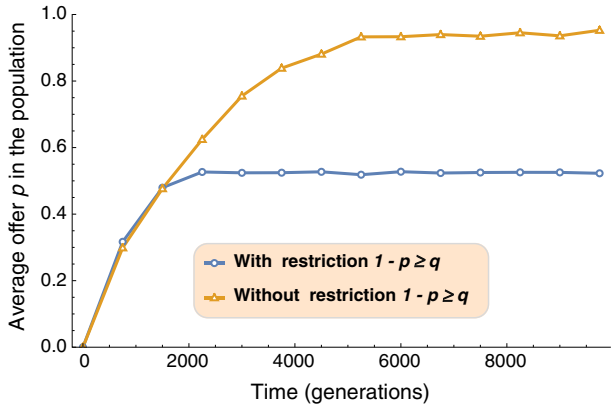

Fig. 1. Evolution of the average ultimatum game offer in the model of Nowak et al. (2000), with and without the restriction $1 - p \geq q$. Only this restriction maintains offers around their fair value of 0.5. Each curve is an average over 20 simulation runs.

that empathy can be represented in models by the assumption $p = q$. We should also be aware of the possibility of a proximal/distal confusion: empathy might be the psychological reason for why people behave fairly but it does not tell us anything about the evolutionary mechanisms that explain empathy (and hence fairness) itself.

3.2.4. The importance of initial conditions

A few papers have reported the evolution of fairness to be dependent on the initial conditions of the model (Chiang, 2008; Roth & Erev, 1993), but few have reported varying their initial settings. A common initial setting is to use random values for individuals' offers and acceptance thresholds. This may seem like a good idea at first, but it presents at least three drawbacks. Biologically speaking, using random values means assuming that some individuals in the population are already fair at time t=0. Noise-based models also show that noise in acceptance thresholds is enough for fair offers to evolve, and using random initial values is precisely a way of introducing noise into the model. Finally, fairness defined as a 0.5 offer has the particularity of corresponding to the average of random values between 0 and 1. This fact needs to be kept in mind, in particular with noise-based models which assume weak selection or high mutation rates, because drift alone will be able to produce offers that look fair when averaged over thousands of generations. A simple way to rule out this interpretation is to provide distributions of offers, which allow to identify if and where modes happen in the distribution.

Hence, we recommend running simulations with at least a ($p = 0$, $q = 0$) initial condition (which corresponds to the most plausible ancestral state, where all individuals are selfish and do not care about fairness), possibly also including ($p = 1$, $q = 1$), ($p = 0.5$, $q = 0.5$), and random initial conditions.

3.2.5. Reciprocity

Reporting the final distributions of offers also helps because there is a trivial way in which fair interactions can evolve: when interactions are repeated, if individuals have equal chances to play the role of proposer or responder, they will on average get a payoff of 50%. When interactions are repeated, it is thus important to report offers instead of mere average payoffs, and distributions rather than averages whenever possible.

3.2.6. Choosing parsimony over realism

There seems to be a trend of producing models showing that fairness can evolve "without": that is, without this or that particular mechanism. Duan and Stanley (2010) want to "reduce the complexity of the rules" of the model (p. 1, paragraph 1), Wang et al. (2014) report the evolution of fairness "even in [an] information-deficiency situation" (p. 5 paragraph 3). Ichinose and Sayama (2014) "propose a new evolutionary model of UG to show that fairness can evolve without additional information such as reputation, empathy, or spatial structure" (p. 2, paragraph 2). Producing simpler models is a good thing because it allows us to identify which conditions are really necessary for the evolution of fairness, and which are irrelevant. At the same time, it will be difficult to make simpler models than noise-based models. Do we need to conclude that human fairness comes from noise because it is the most parsimonious explanation? To us, the priority is not to produce simpler models but to start tackling the evolution of fairness in more realistic situations than the UG (see Section 4.4) or to start comparing the models (see Section 3.2.8). Then only will we know whether it is acceptable to remove reputation or spatial structure from the models, even though it is a well-known empirical fact that humans care a great deal about reputation (Bateson, Nettle, & Roberts, 2006; Haley & Fessler, 2005; Leary & Kowalski, 1990), or that spatial structure characterizes human populations.

It is also important to see that parsimony often comes at a cost regarding the biological credibility of the model. For instance, postulating that a sense of fairness evolved biologically through "noise" or "randomness" is an extremely strong assumption given the costs and centrality of fairness in our daily social life (but this also depends on whether one thinks the models explain the evolution of fairness in the UG only, or the evolution of a *sense* of fairness more generally, see Section 4.4).

3.2.7. Mixing different mechanisms

Some models put different mechanisms into the same model. Wang et al. (2014) investigate the effect of random allocation, but in a spatial population structure. Szolnoki, Perc, and Szabó (2012) investigate the effect of empathy in a spatially structured population. The implication is straightforward: in these cases, it is difficult to distinguish what mechanism really drives the evolution of fairness. We hope that this review will help to avoid these sorts of problems in the future by clearly identifying the different mechanisms that can influence the evolution of fairness.

3.2.8. No comparisons between models

Our final theoretical concern is the rarity of comparisons between models. Although almost all authors cite previous work, virtually none of them state how their model improves our understanding of the origins of human fairness (other than being more parsimonious in the senses described above). An emblematic example comes from Martin Nowak, whose models are included in 4 out of the 6 families identified in Section 2 (Nowak et al., 2000; Page & Nowak, 2002; Page et al., 2000; Rand et al., 2013). Although we should be thankful for his often pioneering work, this profusion of models with very little discussion of their relative strength makes it hard to say which model, if any, should be favored as a candidate to explain human fairness. It is understandable that a large variety of models were produced as the field began, but no new family of models has been produced in the last 10 years. Hence, it may be time to begin comparing models instead of simply continuing to produce incremental variants on previous ones.

As a first step in this direction, we reproduced five of the mainstream models highlighted in Section 2 using agent-based simulations: André and Baumard (2011); Nowak et al. (2000); Page and Nowak (2002); Page et al. (2000); Rand et al. (2013). We will not analyze these simulations here, as a proper analysis would require an article of its own. We only make them available to the community, hoping that some readers will find them useful. The complete models are available online at github.com, figshare.com, and on the first author's personal website (see the Acknowledgments section). They are all coded in Netlogo (Wilensky, 1999), a "low-threshold-no-ceiling" agent-oriented programming language. Netlogo provides an intuitive interface that can be used to explore the parameter space of each model without having to make any changes to the code. Methods for each model are reproduced for convenience in SI section 4. We hope that some readers will find the models useful, but we are aware that a real comparison

cannot be made entirely on theoretical basis. This concern is the subject of Section 4.

4. Links between the theory and the experimental data

4.1. Timescales

A first problem concerns the time-scale of the "evolution" of fairness that is referred to. Authors refer to at least three different timescales:

- an evolutionary, long-term timescale, during which human ancestors could have evolved a biological "sense" of fairness.
- a cultural, medium-term timescale, during which people learn social norms of fairness in their daily life.
- a short-term timescale, limited to the duration of a laboratory experiment, in which people choose their strategies either through careful reasoning or through trial-and-error learning.

Table 1 provides an overview of the timescales studied by each author, but many authors, with a few notable exceptions (Binmore & Samuelson, 1994; Gale et al., 1995), do not explicitly specify the timescale they are studying. Hence, we sometimes had to interpret what authors meant by "evolution", which renders our categorization somewhat subjective; in SI section 1 we provide the exact quotations on which our classification is based.

Many authors argue that their model can be interpreted in terms of both biological and cultural evolution. For example, Rand et al. (2013) state in their abstract that "natural selection favors fairness" and later that they study "the ultimate evolutionary explanation for why we should have come to possess such fairness preferences", which seems to imply that they are investigating a biological device. A few lines later, however, they state that their model "could describe genetic evolution or cultural evolution through social learning" and suggest cultural equivalents for mutations. Their discussion comes back to biology through the use of terms such as "weak selection" and "mutation rate", but the paper ends with a behavioral experiment that should probably be interpreted in terms of cultural evolution.

Some authors think it is a feature of evolutionary models to be interpreted in such different ways. It is true that the dynamics of cultural evolution, biological evolution and even learning can be described with the same equations (Harley, 1981). Remaining vague about the intended timescale allows the model to be more general and also more consensual for the irritable reviewers. Nonetheless, other authors (us included) think that this refusal to specify timescale is one of the elements that has negatively impacted our understanding of the origins of human fairness in the last years. When the first models of the evolution of fairness came out thirty years ago, it could be rightfully argued that there was not enough experimental evidence to make up one's mind. Today this is less and less true. An impressively ample literature has been developed on the developmental trajectory of fairness in children (Fehr, Bernhard, & Rockenbach, 2008; Geraci & Surian, 2011; Schmidt & Sommerville, 2011; Sloane, Baillargeon, & Premack, 2012; Warneken, Lohse, Melis, & Tomasello, 2011), its neurological basis (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Sanfey, Rilling, & Aronson, 2003; Tabibnia, Satpute, & Lieberman, 2008), its similarities with other species (Bräuer & Hanus, 2012; Brosnan & de Waal, 2014; Warneken & Tomasello, 2009), and, to a lesser extent, its universality (Henrich, 2004; Marshall, Swift, Routh, & Burgoyne, 1999). Although the degree to which fairness is cultural or biological remains an open question, and we do not suggest that authors take up a stand exclusively on one or the other side, these works are important for theorists because they should help them to make appropriate assumptions when building their models.

4.2. Is the UG just a pretext?

An important problem somewhat related to the issue of timescales is that in many cases, authors who use the same terms (fairness, ultimatum game, inequity aversion…) are actually trying to explain different things. For example, some authors are trying to understand the origin of the variability of decisions in the UG (i.e., individual-level behavior). As such, they try to explain why the modal offer in the empirical data is usually between 40% and 50%, and why other offers are distributed between 0% and 40%. In this view, offers in the UG are an object of study per se, and this study does not necessarily require a "theory of fairness" in the sense of what the "purpose" of fairness in our daily life is, be it evolutionary or cultural.

Other authors, on the contrary, do not take models of the evolution of fairness at face value, implying that UG decisions are the kind of things that can evolve. Rather, their interpretation is that psychological mechanisms that give rise to fair decisions in the UG can evolve. Those authors are thus perfectly satisfied to find that their model only predicts offers of exactly 50%, even though this contradicts the empirical data, since they are using the UG not as an object of study per se but as a convenient way to model an asymmetric power struggle between two individuals. In this sense, the UG is more to be compared with the classical "Hawk and Dove" or "war of attrition" games used in the animal literature on asymmetric contests (Hammerstein, 1981; Maynard Smith & Parker, 1976). The evolution of equal offers in these models is usually meant to represent the long-term evolution of a "sense" or "taste" for fairness in humans, not the dynamics of offers observed in behavioral experiments.

This last interpretation explains why criticisms such as "models based on reputation or repeated interactions can not explain fairness in the empirical UG because the empirical UG is one-shot and anonymous" are misguided. These models predict that reputation or repeated interactions outside the lab (cultural explanation) or at the ultimate level (biological explanation) have led to the evolution of a sense of fairness which now functions more or less automatically: it produces the kind of behaviors we observe in the UG even when reputation or repeated interactions are absent. Another way to put it is to say that those models suppose that fairness is suboptimal in one-shot anonymous economic games but optimal in a wider framework including reputation or repeated games, to the point where fairness could have been biologically "hardwired" or have become a social norm.

For improved clarity, we suggest that authors specify whether they consider the UG as an object of study per se or only as a convenient way to model a bargaining problem in the larger framework of the evolution of a sense of fairness in humans.

4.3. Human specificity

The problem of resource division is an important problem in evolutionary biology. As such, it has already been investigated outside the human context. Models of reproductive skew, for example, try to understand why reproduction is more or less equally shared in some species but more biased towards a few dominant individuals in other species (Johnstone, 2000; Vehrencamp, 1983). These models are based on the same mechanism as partner choice-based models of fairness, in that an individual's outside options determine the division of the resource. Models of biological markets investigate how supply and demand affect the price at which a commodity is exchanged between two classes of traders (Noë & Hammerstein, 1994; Noë, Schaik, & Hooff, 1991). Models of asymmetric contests deal with the division of a resource when individuals differ in terms of competitive power (Hammerstein & Parker, 1982; Maynard Smith & Parker, 1976). Spite and spatial structure are two other mechanisms that have been widely investigated outside a human context (Gardner & West, 2004; Lehmann et al., 2006).

There is every reason for these models to be a great source of inspiration for human-related modeling, but it is relatively rare to see them cited in the human literature. More importantly, comparing fairness models to non-human models is the occasion to address the question of human specificity. Although the empirical difference between humans and other species' social skills is still a matter of great debate,

most scholars agree that there is something special about human fairness. Almost all articles on modeling the evolution of fairness feature in their introduction a reminder of the extraordinary human capacity to care about the interests of – even unrelated – others. Unfortunately, almost none return to this point in the discussion in order to assess how their model helps to explain this specificity. After all, alternating roles, noise, spite, and spatial population structure are not restricted to human ecologies, so why should fairness have evolved only – or mainly – in humans?

Out of all the mainstream models we reviewed, only one explicitly addresses the question of human specificity (see Table 1). We thus suggest that authors specify the peculiarities of human ecology, culture, or brain that they believe have allowed fairness to develop in humans more than in any other species. The hypotheses can be purely speculative, but it should be possible to test them empirically. In turn, the empirical test can serve as a way to evaluate the models' biological plausibility and provide a starting point for cross-model comparisons, two things that, although they are obviously necessary, are seldom available at present.

4.4. Does the model explain more than equal divisions?

The focus of this paper has been on fairness *in the UG*, as a synonym for equal divisions of money. To our knowledge, virtually all models of the evolution of fairness are about the evolution of such equal divisions. But what can these models say about the evolution of fairness outside the UG? Fairness in our daily life indeed consists of more than just equal divisions. For instance, work on equity theory, the behavioral theory of distributive justice, has demonstrated that people strongly prefer divisions that are matched to contributions: the more someone contributes, the more they should receive in return (Adams, 1963). Although there is no debate that these types of "meritocratic" preferences constitute an important aspect of fairness, we are unaware of any models that have attempted to model the evolution of such preferences. Hence, an obvious way to start comparing the six mechanisms identified in Section 2 is to investigate whether each mechanism, on top of being able to explain the evolution of equal divisions, can also explain the evolution of other aspects of human fairness such as meritocratic divisions. Fairness can also characterize the right amount of effort to invest into cooperation, or the right amount of punishment to give to someone (for a review of models of the evolution of cooperation, see Lehmann and Keller (2006)). Investigating whether models can explain the fair behaviors we observe in such situations seems a promising avenue of research.

4.5. Is the UG meaningful for the study of fairness?

Many authors have questioned the assumption that the UG constitutes a good empirical measure of fairness, or a measure of fairness at all. Some have argued that equal divisions only reflect proposers' fear that their offers will be rejected, and indeed when responders are not allowed to reject offers (as is the case in the game called the dictator game), the modal offer is much lower (Camerer, 2003). In the same vein, some authors have suggested that punishment has been a central force in the evolution of human fair or cooperative behaviors (Fehr & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003). Other scholars (Baumard & Sperber, 2010;Ceci, Kahan, & Braman, 2010) have pointed out that the lack of information provided in the UG requires subjects to answer many questions by themselves: where does the money come from? Is there a right for the proposer to keep it because the experimenter gave it to her? Does the UG represent a competitive or cooperative real-life interaction? Hence, special interpretations of the game could explain special behaviors. Kirchsteiger (1994) has even suggested that envy on the side of responders (and not fairness) could be responsible for the observed rejections.

The question is thus to know whether equal offers, commonly referred to as "fair" offers in the literature, are the product of a sense of fairness *at the psychological level*. We need to be clear that theoretical models do not really help to shed light on such proximate mechanisms. Any model showing why it is advantageous to refuse small offers can always be implemented psychologically in two very different ways: through the existence of a genuine sense of fairness, or through the existence of preferences for revenge/punishment/etc. This is a question that will have to be settled empirically and is beyond the scope of this paper. If authors are inclined to think that there is no empirical evidence for the existence of a sense of fairness, then they will interpret the models as explaining behaviors only, and the "fair" label given to these behaviors as a label that does not reflect the existence of a fair psychology. But other authors will argue that if it is very likely that some of the equal offers we observe in UGs come from selfish strategic reasoning, even in the dictator game many people make non-zero offers (around 60% according to a meta-analysis by Engel (2011)). Better yet, Engel (2011) shows that our grim view of dictator games might be due to an over-emphasis on student populations: in middle age populations, 50% of people give exactly 50% of the money to their partner. These decisions cannot be explained by selfish strategic reasoning. Hence, no matter whether decisions *in the UG* come from "a sense of fairness" or not at the psychological level, some might argue that the existence of a genuine sense of fairness is plausible based on other data or real-life situations, and its evolution needs to be explained.

But why use the UG in this case and not the dictator game to model the evolution of fairness? Although we would welcome such models based on the dictator game, as explained in Section 4.2 the UG is often used as a pretext to model an asymmetry of bargaining power between two individuals. In this perspective, the UG is used to investigate the evolution of psychological mechanisms which produce equal offers, not the evolution of equal offers directly. Hence, because in the absence of particular mechanisms natural selection favors selfishness in the UG, using this game as a basis to understand how fair behaviors can evolve theoretically is not misguided.

5. Conclusion

More than thirty years after the first clear experimental evidence of fairness in humans, it is heartening to see that there is no shortage of theoretical explanations for its paradoxical existence. Scholars from many different fields have put forward a wide variety of hypotheses, promising a rich debate in years to come. We hope that this review will contribute to the debate by providing an initial classification of the competing theories and by clarifying the mechanisms at play in each theory. Although the field is not without its problems, none of them is insurmountable. Our main recommendation is to create closer links between the models and real-world data by explicitly specifying: (1) the proposed timescale of the evolution of fairness; (2) the assumed function and importance of fairness in daily human life; (3) how the model helps understand the human specificity of fairness; and (4) whether the model can explain more than the evolution of equal divisions. We hope that some researchers will find these guidelines helpful and that they will encourage others to continue on with the comparative work that we have started in this review.

Acknowledgments

We thank Stuart A. West, Andy Gardner and an anonymous reviewer for valuable comments on the manuscript.

SD thanks the *Région Ile-de-France* for funding this research through a 2012 DIM *"Problématiques transversales aux systèmes complexes"* grant, and thanks the Institut des systèmes complexes and the doctoral school *"Interdisciplinaire Européenne Frontières du Vivant ED 474, Programme doctoral Bettencourt, Universities Paris Descartes and Paris Diderot"* for

their support. This work was supported by ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

The data and code for the simulations presented in this paper are archived online on the first author's website http://stephanedebove.net/?p=239 and on Figshare.com https://figshare.com/s/651cb8eab2a5288b9c6c. The source code for the 5 replicated models can also be downloaded or forked on Github.com : https://github.com/BigNoob/fairness-netlogo/. It is also available on request from the first author. The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2016.01.001.

References

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal Psychology*, *67*, 422–436.

Alexander, J. (2007). *The structural evolution of morality*. Cambridge University Press.

André, J. -B., & Baumard, N. (2011). Social opportunities and the evolution of fairness. *Journal of Theoretical Biology*, *289*, 128–135.

Barclay, P., & Stoller, B. (2014). Local competition sparks concerns for fairness in the ultimatum game. *Biology Letters*, *10*, 1–4.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*, 412–414.

Baumard, N., & Sperber, D. (2010). Weird people, yes, but also weird experiments (Commentary on: Joseph Henrich, Steven J. Heine, Ara Norenzayan (2010) The weirdest people in the world?). *Behavioral and Brain Sciences*, *33*, 80–81.

Bethwaite, J., & Tompkinson, P. (1996). The ultimatum game and non-selfish utility functions. *Journal of Economic Psychology*, *17*, 259–271.

Binmore, K. (2005). *Natural justice*. Oxford University Press.

Binmore, K., & Samuelson, L. (1994). An economist's perspective on the evolution of norms. *Journal of Institutional and Theoretical Economics*, *150*, 45–63.

Bräuer, J., & Hanus, D. (2012). Fairness in non-human primates? *Social Justice Research*, *25*, 256–276.

Brosnan, S., & de Waal, F. (2014). Evolution of responses to (un) fairness. *Science*, *346*, 1–10.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*, *vol. 32*, Princeton, New Jersey: Princeton University Press.

Ceci, S. J., Kahan, D. M., & Braman, D. (2010). The WEIRD are even weirder than you think: Diversifying contexts is as important as diversifying samples. *Behavioral and Brain Sciences*, *33*, 27–28.

Chiang, Y. -S. (2007). The evolution of fairness in the ultimatum game. *The Journal of Mathematical Sociology*, *31*, 175–186.

Chiang, Y. -S. (2008). A path toward fairness: Preferential association and the evolution of strategies in the ultimatum game. *Rationality and Society*, *20*, 173–201.

da Silva, R., Kellermann, G. A., & Lamb, L. C. (2009). Statistical fluctuations in population bargaining in the ultimatum game: Static and evolutionary aspects. *Journal of Theoretical Biology*, *258*, 208–218.

Debove, S., Baumard, N., & André, J. -B. (2015). Evolution of equal division among unequal partners. *Evolution*, *69*, 561–569.

Duan, W. -Q., & Stanley, H. E. (2010). Fairness emergence from zero-intelligence agents. *Physical Review E*, *81*, 026104.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*, 583–610.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*, 1079–1084.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*, 817–868.

Forber, P., & Smead, R. (2014). The evolution of fairness through spite. *Proceedings of the Royal Society B*, 281.

Gale, J., Binmore, K., & Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, *8*, 56–90.

Gardner, A., & West, S. A. (2004). Spite and the scale of competition. *Journal of Evolutionary Biology*, *17*, 1195–1203.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science*, *14*, 1012–1020.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*, 153–172.

Güth, W., & Kocher, M. (2013). *More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature*.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367–388.

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching? *Evolution and Human Behavior*, *26*, 245–256.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I & II. *Journal of Theoretical Biology*, *7*, 17–52.

Hammerstein, P. (1981). The role of asymmetries in animal contests. *Animal Behaviour*, 193–205.

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology*, *96*, 647–682.

Harley, C. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology*, 611–633.

Harms, W. (1997). Evolution and ultimatum bargaining. *Theory and Decision*, 147–175.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*, 3–35.

Hoel, M. (1987). Bargaining games with a random sequence of who makes the offers. *Economics Letters*, *24*, 5–9.

Huck, S., & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, *28*, 13–24.

Ichinose, G., & Sayama, H. (2014). Evolution of fairness in the not quite ultimatum game. *Scientific Reports*, *4*, 5104.

Iranzo, J., Flora, L. M., Moreno, Y., & Sánchez, A. (2012). Empathy emerges spontaneously in the ultimatum game: Small groups and networks. *PloS One*, *7*, e43781.

Iranzo, J., Román, J., & Sánchez, A. (2011). The spatial ultimatum game revisited. *Journal of Theoretical Biology*, *278*, 1–10.

Johnstone, R. A. (2000). Models of reproductive skew: A review and synthesis. *Ethology*, *106*, 5–26.

Killingback, T., & Studer, E. (2001). Spatial ultimatum games, collaborations and the evolution of fairness. *Proceedings. Biological sciences / The Royal Society*, *268*, 1797–1801.

Kirchsteiger, G. (1994). The role of envy in ultimatum games. *Journal of Economic Behavior & Organization*, 2681.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science (New York, N.Y.)*, *314*, 829–832.

Leary, M., & Kowalski, R. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*, 34–47.

Lehmann, L., Bargum, K., & Reuter, M. (2006). An evolutionary analysis of the relationship between spite and altruism. *Journal of Evolutionary Biology*, *19*, 1507–1516.

Lehmann, L., & Keller, L. (2006). The evolution of cooperation and altruism — A general framework and a classification of models. *Journal of Evolutionary Biology*, *19*, 1365–1376.

Marshall, G., Swift, A., Routh, D., & Burgoyne, C. (1999). What is and what ought to be popular beliefs about distributive justice in thirteen countries. *European Sociological Review*, *15*, 349–367.

Maynard Smith, J., & Parker, G. (1976). The logic of asymmetric contests. *Animal Behaviour*, 159–175.

Maynard Smith, J., & Price, G. (1973). The logic of animal conflict. *Nature*, *246*.

Nash, J. (1950). The bargaining problem. *Econometrica*, *18*, 155–162.

Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*, 1–11.

Noë, R., Schaik, C., & Hooff, J. (1991). The market effect: An explanation for pay-off asymmetries among collaborating animals. *Ethology*, *87*, 97–118.

Nowak, M. a., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*, 1773–1775.

Page, K. M., & Nowak, M. A. (2002). Empathy leads to fairness. *Bulletin of Mathematical Biology*, 1101–1116.

Page, K. M., & Nowak, M. A. (2001). A generalized adaptive dynamics framework can describe the evolutionary ultimatum game. *Journal of Theoretical Biology*, *209*, 173–179.

Page, K. M., Nowak, M. A., & Sigmund, K. (2000). The spatial ultimatum game. *Proceedings. Biological sciences / The Royal Society*, *267*, 2177–2182.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., & Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 2581–2586.

Roth, A., & Erev, I. (1993). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, *8*, 164–212.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, *50*, 97–110.

Sánchez, A., & Cuesta, J. A. (2005). Altruism may arise from individual selection. *Journal of Theoretical Biology*, *235*, 233–240.

Sanfey, A., Rilling, J., & Aronson, J. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–1758.

Schmidt, M. F. H., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PloS One*, 6.

Skyrms, B. (1996). *Evolution of the social contract*.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, *23*, 196–204.

Stahl, I. (1977). An n-person bargaining game in the extensive form. *Mathematical economics and game theory*, *vol. 1*, .

Szolnoki, A., Perc, M., & Szabó, G. (2012). Defense mechanisms of empathetic players in the spatial ultimatum game. *Physical Review Letters*, *109*, 078701.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness. *Psychological Science*, *19*, 339–347.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Vehrencamp, S. L. (1983). Optimal degree of skew in cooperative societies. *American Zoologist*, *23*, 327–335.

Wang, X., Chen, X., & Wang, L. (2014). Random allocation of pies promotes the evolution of fairness in the ultimatum game. *Scientific Reports*, *4*, 4534.

Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science*, *22*, 267–273.

Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences, 13*, 397–402.
West, S. A., & Gardner, A. (2010). Altruism, spite, and greenbeards. *Science (New York, N.Y.), 327*, 1341–1344.
West, S. a., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*, 415–432.
West, S. S. A., Mouden, C. E., Gardner, A., & El Mouden, C. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior, 32*, 231–262.
Wilensky, U. (1999). *NetLogo*.
Zollman, K. J. (2008). Explaining fairness in complex environments. *Politics, Philosophy & Economics, 7*, 81–97.