PARIS DESCARTES UNIVERSITY

Doctoral School Frontières du Vivant

PhD Thesis in Evolutionary Psychology

The evolutionary origins of human fairness

By Stéphane DEBOVE

École normale supérieure
Institut de Biologie de l'ENS & Institut Jean Nicod

PhD Advisors

Jean-Baptiste ANDRÉ Nicolas BAUMARD

Presented and defended publicly on October 29th, 2015

Jury

Pr. Redouan BSHARY	Rapporteur - Université de Neuchâtel
Pr. Ronald NOE	Rapporteur - Université de Strasbourg
Pr. Pat BARCLAY	Examinateur - University of Guelph
Pr. Michel RAYMOND	Examinateur - Université Montpellier II
Dr. Jean-Baptiste ANDRÉ	PhD Advisor - Université Montpellier II
Dr. Nicolas BAUMARD	PhD Advisor - École normale supérieure
Pr. Régis FERRIÈRE	PhD Advisor - Ecole normale supérieure

LES ORIGINES ÉVOLUTIONNAIRES DU SENS DE L'ÉQUITÉ CHEZ L'HOMME

SOUTENANCE DE THÈSE
STÉPHANE DEBOVE

LE 29 OCT.2015
À 14 HEURES
ENS, SALLE DES ACTES

Abstract:

Humans care about fairness and are ready to suffer financial losses for the sake of it. The existence of such costly preferences for fairness constitutes an evolutionary puzzle. Recently, some authors have argued that human fairness can be understood as a psychological adaptation evolved to solve the problem of sharing the costs and benefits of cooperation. When people can choose with whom they want to cooperate, sharing the costs and benefits in an impartial way helps to be chosen as a partner and brings direct fitness benefits. In this theory, partner choice is thus the central mechanism allowing the evolution of fairness. Here, we offer an interdisciplinary study of fairness to put this theory to the test. After a review of competing theories (Paper 1, in review), we build game-theoretical models and agent-based simulations to investigate whether partner choice can explain two key aspects of human fairness: the wrongness to take advantage of one's strength to exploit weaker people (Paper 2, *Evolution*), and the appeal of distributions where the reward is proportional to the contribution (Paper 3, in review). We show that partner choice succeeds at explaining these two characteristics. We also go towards more realistic and mechanism-oriented simulations by trying to evolve fair robots controlled by simple neural networks. We then test the theory empirically, and show that partner choice creates fairness in a behavioral experiment (Paper 4, *Proceedings of the Royal Society B*). We develop a collaborative video game to assess the cross-cultural variation of fairness in distributive situations, and present results coming from a Western sample (Paper 5, in preparation). We review the experiments looking for fairness in non-human animals, and discuss why fairness would have been more prone to evolve in humans than in any other species, despite partner choice being an evolutionary mechanism far from restricted to the human species. Finally, we discuss three common misunderstandings about the partner choice theory and identify interesting directions for future research.

Keywords : human fairness, equity, justice, partner choice, biological markets, cooperation, morality

Titre : Les origines évolutionnaires du sens de l'équité chez l'Homme

Résumé :

L'Homme attache de l'importance à l'équité et est prêt à aller jusqu'à subir des pertes financières pour la défense de l'équité. Cet attachement coûteux à l'équité constitue un paradoxe pour les théories de l'évolution. Récemment, certains auteurs ont proposé de voir le sens de l'équité comme une adaptation psychologique évoluée pour résoudre le problème du partage des coûts et bénéfices de la coopération. Quand il est possible de choisir avec qui coopérer, partager les coûts et bénéfices d'une manière impartiale aide à être choisi comme partenaire social et procure des bénéfices directs en terme de valeur sélective. Dans cette théorie, le choix du partenaire est donc le mécanisme central permettant l'évolution du sens de l'équité. Ici, nous proposons une étude interdisciplinaire de l'équité pour mettre cette théorie à l'épreuve. Après une revue des théories en compétition pour expliquer l'équité (Article 1, en cours de revue), nous développons des modèles de théorie des jeux et des simulations individu-centrées pour savoir si le choix du partenaire permet d'expliquer deux éléments-clés de l'équité: le refus de profiter de sa force pour exploiter les plus faibles (Article 2, *Evolution*), et l'attrait des distributions dans lesquelles la rétribution est proportionnelle à la contribution (Article 3, en cours de revue). Nous montrons que le choix du partenaire permet d'expliquer ces deux caractéristiques. Nous produisons également des simulations plus réalistes et prenant mieux en compte les mécanismes d'évolution en essayant de faire évoluer des robots qui se comportent de manière équitable. Nous testons ensuite la théorie de façon empirique, et montrons que le choix du partenaire crée des distributions équitables dans une expérience comportementale (Article 4, *Proceedings of the Royal Society B*). Nous développons un jeu vidéo collaboratif pour estimer l'importance de la variabilité interculturelle de l'équité dans des situations de justice distributive, et présentons des résultats obtenus sur un échantillon de sujets occidentaux (Article 5, en préparation). Nous passons en revue les expériences cherchant de l'équité chez les animaux non-humains, et discutons pourquoi un sens de l'équité aurait eu plus de chances de se développer chez l'Homme que dans une autre espèce, alors que le choix du partenaire est loin d'être un mécanisme évolutionnaire restreint à l'Homme. Enfin, nous discutons trois malentendus classiques sur la théorie du choix du partenaire et identifions des directions de recherche intéressantes pour le futur.

Mots-clés : équité, justice, choix du partenaire, marché biologique, coopération, morale.

To Sophie.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the "Région Ile de France" for funding my PhD through a "2012 Domaine d'Intérêt Majeur, Problématiques transversales aux systèmes complexes", and the "Institut des systèmes complexes" (ISC) for selecting my research project.

Thank you to the biological market superheroes Pat Barclay, Redouan Bshary, Ronald Noe and Michel Raymond for agreeing to be part of my PhD jury. I look forward to our discussions on the topic.

Thank you to my doctoral school, "Frontières du Vivant", and the "Centre de Recherches Interdisciplinaires", for providing intellectual, social and financial support. It is not exaggerated to say that my PhD experience would have been much poorer without them. The people I met there, research teaching I did, and courses I attended will not be forgotten.

Thank you to my two teams, the Evolution and Social Cognition team and the Eco-Evolutionary Mathematics team, for scientific discussions around beers or not. I found especially nice to hang out with people without fearing to be kicked for saying that human behavior is partly biologically determined.

Thank you to my family and in particular my parents for providing me with such a stimulating environment in my childhood, and never interfering in my career choices. I am not sure they really had the choice, but thank you anyway. And thank you to la Ces for this wonderful defence poster!

I am greatly indebted to my advisors Jean-Baptiste and Nicolas. Unlike a non-negligible number of students I know, I strongly enjoyed my PhD and will have no nightmare stories to tell my children. No doubt they are largely to be blamed for this. I feel lucky to have been able to work on such a fascinating subject in an interdisciplinary framework, at a time when evolutionary psychology is still a dirty word in France. I also feel lucky to have been able to benefit from their earlier work on the evolution of fairness - without them clearing the way, my thesis would be much shorter and shallower. I will remember conversations in which I was useless as important intellectual lessons, but I will also remember their human qualities. I am in particular thankful for their great availability and the freedom they gave me to conduct my research. Who could really complain when allowed to build video games during his PhD? All this to say that if they need a recommendation letter for their next PhD student I will be happy to write one, and that I hope this thesis will not mark the end of our collaboration but rather the beginning of an hopefully long one.

The following actors have nothing in common but are mentioned because the effort to imagine what my life or research would look like without them is too important. They are the old post-bac french scholarship system that has now been discontinued, the Ecole normale supérieure and its rich interdisciplinary environment, the open-source movement, and more generally all the scientists I'm standing on the shoulders of, with or without knowing it.

Last but not least, thank you to Ysé, for tolerance at my countless hours on PC, patience at my running jokes that can run for years, and support for engaging in this activity that I like to call research and that she likes to call useless.

Part I

Introduction

Fair trade sales generated more than €944 millions in producers revenues in 2012-2013, a 113 percent increase compared to 2008 (Fairtrade International, 2014). These figures show that quite a large number of people are willing to pay more for their chocolate bar only to be sure that a stranger on the other side of the world is treated fairly. To me, this is quite remarkable. The non-negligible sums of money given in pay-what-you-want pricing systems, in which people can freely and anonymously decide how much to pay for a digital download, is yet another remarkable manifestation of fairness in the modern world (Kim et al., 2009). But fairness is not limited to market exchanges. Seeing someone skip the line, park on a space for disabled people, or win a football match with a hand goal are unfair situations to which we commonly react. The Google Suggestion tool is excellent for giving us an overview of the variety of situations in which people express fairness concerns. Among the most common search queries beginning by *"is it fair to"* can be found *"is it fair to keep my cat indoors", "is it fair to have extramarital affair", "is it fair on the part of humans to rear sheep", "is it fair to have a baby at 40", "is it fair to ask boyfriend to stop drinking", "is it fair to have an abortion"*, or *"is it fair to use legendary pokemon"*... Fairness with no doubt has a strong impact on human lives, plays an important role in human psychology, and is taken into account when making many futile as well as life-changing decisions.

For an evolutionary biologist, the existence of those preferences for fairness is puzzling. As well illustrated by the fair trade example, preferences for fairness are costly: people are ready to incur costs for the sake of fairness. This behavior is not readily explained by evolutionary theories, which predict that only behaviors providing fitness benefits should evolve. A question that has generated a lot of debates in the last decades is thus: how could natural selection explain the evolution of such costly preferences for fairness?

Quite recently, Baumard et al. (2013) have suggested to see fairness as a psychological adaptation evolved by social selection. In this perspective, fairness is natural selection's solution to the problem of how to share the costs and benefits of cooperation in an environment where people can choose who they want to cooperate with. It is easy to see that there is an intrinsic tension in this situation: from an evolutionary perspective, people should want to both keep most of the benefits for themselves, but also remain able to attract social partners to cooperate again in the future. Baumard et al. (2013) propose that fairness solves this exact problem, by providing the best compromise between over-generosity and over-selfishness. Preferring fair divisions makes people maximise their fitness, because those divisions balance the costs and benefits of being too generous or too stingy (see Chapter 2 for a better explanation of the theory). Because selfishness is overriden by the necessity

to be chosen as a social partner, "partner choice" can be identified as the central evolutionary force driving the evolution of fairness in this theory.

The objective of my PhD was to put this theory to the test and investigate the extent of its explanatory power, using both theoretical and empirical methods. But my work also has a wider scientific context of course. As I will have the occasion to come back to it several times, to avoid overloading the introduction I will only present the general context here.

The context of my work is to introduce more partner choice to the study of the evolution of cooperation[1]. Before the 90's, classical evolutionary models typically relied on reciprocity to explain the evolution of cooperation among non-kins (Trivers, 1971). Hence, they were considering repeated interactions taking place between partners paired at random and having to make a binary choice between (i) cooperating with each other or (ii) not-cooperating at all. In other words, models neglected the possibility that some individuals might prefer to cooperate with other partners than the ones they were currently paired with. This assumption is problematic both empirically and theoretically.

Empirically, it neglects the fact that animals often have a variety of partners at disposition to cooperate with. Hence, some authors have spoken in favor of integrating more partner choice to the study of cooperation, as had been done successfully in the study of sexual selection or mutualism (Noë et al., 1991; Noë and Hammerstein, 1995; Roberts, 1998). When partner choice is considered, social life looks like a "market" in which individuals compete to be recruited in cooperative ventures. Many models on partner choice and cooperation have been produced since then (Aktipis, 2004; McNamara et al., 2008; Nesse, 2007; Johnstone and Bshary, 2008; Barclay, 2011) and a biological market perspective has often helped to explain the empirical data (Bshary and Schäffer, 2002; Fruteau et al., 2009; Schwagmeyer, 2014), showing the relevance of the approach. But if this approach has often been used to study how much to cooperate or how much to help (i.e. how much to *invest* into cooperation), it has less often been used to study how to *divide* the benefits of cooperation once they have been produced (but see Chapter 3 for a discussion of reproductive skew models). Compared to the vast literature on the evolution of cooperation, an originality of the present thesis is to investigate the consequences of partner choice on the *division* of a resource (but see sections 5.5 and 10.1 for why this distinction is not entirely justified).

[1]Because the field is famous for being a terminological mess, I provide a glossary in Table 1. My definitions are those of West et al. (2011).

Term	Definition
Actor	The focal individual performing a behaviour.
Altruism	A behaviour that is costly to the actor and beneficial to the recipient or recipients. Costs and benefits are defined on the basis of the lifetime direct fitness consequences of a behaviour.
Cooperation	A behaviour that provides a benefit to another individual (recipient), and the evolution of which has been dependent on its beneficial effect for the recipient.
Kin selection	Process by which traits are favoured because of their effects on the fitness of related individuals
Mutually-beneficial	A behaviour that is beneficial to both the actor and the recipient.
Recipient	An individual who is affected by the behaviour of the focal actor.
Relatedness	A measure of the genetic similarity of two individuals, relative to the average; the least- squares linear regression of the recipient's genetic breeding value for a trait on the breeding value of the actor
Selfishness	A behaviour which is beneficial to the actor and costly to the recipient.
Social behaviours	Behaviours which have a fitness consequence for both the individual that performs the behaviour (actor) and another individual (recipient).
Spite	A behaviour that is costly to both the actor and the recipient.

Table 1: Glossary (from West et al., 2011). Of particular note is that my use of "cooperation" will encompass both altruistic and mutually-beneficial behaviors. Also, if the term "mutualistic" has been used previously to present the theory of the evolution of fairness by partner choice (Baumard et al., 2013), I will use the term "mutually-beneficial" here instead to prevent confusion with the inter-species behaviors (West et al., 2007).

The second reason why it is problematic not to consider partner choice is that there are good theoretical reasons to think traditional models of cooperation are unable to explain fairness. Because outside options are poor in models without partner choice (if individuals do not like their partner, they will not cooperate at all), individuals are better off accepting almost any form of cooperative interaction, as long as they gain more than what they would gain on their own. As a consequence, any level of investment into cooperation can be selected. The equilibrium is highly indeterminate (Boyd, 2006), and the specific form of cooperation can not be predicted. This result is what economists have called the 'folk theorem' (Fudenberg and Maskin, 1986). It is a problem for our matter because fair cooperation is precisely a specific form of cooperation, in which the costs and benefits are distributed impartially. Hence, it is difficult to see how fairness could be explained by models relying only on partner control.

At this point some people would probably like to get a definition of what I call fairness. I will refrain from doing so, because finding a definition of fairness is precisely one of the things that an evolutionary study like ours allows to do. If we succeed in finding the reason why fairness evolved, we should be able to give a definition of the phenomenon at least at the ultimate level. Nonetheless, a large part of this thesis will refer to fairness in a very specific and restricted sense: an equal division of resources between two individuals. This is the most common definition of fairness in the current evolutionary literature, as a consequence of the desire to explain behavioral experiments in which two people share a sum of money equally for the sake of fairness (Güth et al., 1982). This definition should be challenged though if we wish to explain the real-world data, and I try to do so in Chapters 4, 5, and 10 notably.

In Part II, I start by reviewing the different explanations for the evolution of fairness that exist in the literature. I present in more details the theory of the evolution of fairness by partner choice. I then investigate the explanatory power of this theory in theoretical (Part III) and empirical (Part IV) studies. In Part III, I build evolutionary models to determine whether partner choice can explain two central characteristics of fairness: the wrongness to take advantage of one's strength to exploit weaker individuals, and the appeal of "meritocratic distribution" that obey the rule "reward according to contribution". In Part IV, I test the effect of partner choice experimentally and go on the chimeric quest to find cross-cultural invariants of human fairness. In the discussion (Part V), I try to solve the paradox of a sense of fairness that is often presented as human-specific but would come from an evolutionary mechanism that is far from being limited to humans. I also discuss

three common misunderstandings about the theory and the directions for future research I find the most interesting.

Each chapter presenting results starts with a section "Objectives and summary" providing some context on the work and summarizing the results in one sentence or two. I wrote those sections with the reader in a hurry in mind, but also the reader already familiar with some of the aspects of this thesis, as some chapters are a simple copy of already published articles or articles currently in review. Regarding contributions, unless otherwise noted at the beginning of the chapters, I am responsible for all of the work presented in this thesis (I designed all projects together with my advisors who supervised the work all along, and I happily stole their ideas for writing introductions and conclusions, but I was the only one involved in the actual work of building models, simulations, experiments, collecting and analysing data, and writing papers.).

Part II

Explaining human fairness

Chapter 1

Competing explanations for the evolution of fairness (Paper 1)

"I shared half of the points. It is the right thing to do. It is fair and equal."

A3RT89INMIE

1.1 Objectives and summary

Different explanations for the evolution of fairness coexist. This chapter aims at reviewing and classifying them. I review 36 theoretical models of the evolution of human fairness and identify 6 categories into which they can all be broadly classified: alternating role-based models, reputation-based models, noise-based models, spite-based models, spatial population structure models and empathy-based models. This variety shows that there is currently no consensus in the scientific community as to where fairness could come from. There is also no consensus as to what type of evolution led to fairness: biological, cultural or learning-based (see Table 2 page 26 for a quick overview of this diversity). I identify problems undermining progress in the field: terminology problems, with authors using the same words to mean different things, or modelling problems, with hypotheses getting disconnected from the biological reality. I also emphasize the need to start studying the evolution of fairness outside the ultimatum game, in a wider range of biological situations. For instance, can models explain more than equal divisions of benefits? Can they also explain equitable divisions of benefits, where the output for each individual is made proportional to her input? Although many authors like to start their article with a reminder of the paradoxical existence of fairness *in humans*, very few discuss why their model should be able to explain this human specificity. In Chapters 3 and 4 and in Part V, I will try myself to take these criticisms into account.

The rest of this chapter comes from a paper currently under review at *Evolution and Human Behavior*.

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

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 modelling of the evolution of fairness in a wider and more realistic range of situations.

1.2 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 et al. (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 et al., 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, 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 analyse 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 1.5.4. 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 Kirchsteiger 1994; Bethwaite and

Tompkinson 1996; Fehr and Schmidt 1999). 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 et al. (2009), and see SM1 section 2 (SM, 2015)). Readers interested in these modelling approaches and historical models of bargaining more generally can consult the books by Skyrms (1996), Binmore (2005), 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 a 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 1.4.2) 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 and 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 $p = \epsilon$ and $q = \epsilon$ (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.

Table 2: 17 models of the evolution of fairness and their main characteristics. Classification can be subjective as authors sometimes do not make their interpretation explicit, but SM1 Table 1 (SM, 2015) 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 1.5.2).

	Mechanism	Timescale of evolution	Human-specificity	Restricted to UG
Alexander (2007)	Spatial population structure	Cultural	N/S	No
André and Baumard (2011a)	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)	Requests 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

1.3 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 SM1 section 3.2 (SM, 2015). 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 2 summarizes the classification. SM1 section 1 (SM, 2015) presents the classification of more recent models that we could not include here for reasons of space.

1.3.1 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 et al., 1995; Skyrms, 1996), or have an evolved, biological sense of fairness that they bring and use in the lab.

1.3.2 Reputation-based models

Nowak et al. (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 1.4.2.

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 (2011a) 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 (2011a) 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 modelling, 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 et al. (2015b)).

1.3.3 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 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 et al. (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 an 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 1.5.4 on this point).

1.3.4 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 (West and Gardner, 2010; Lehmann et al., 2006). 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

$$\frac{2}{2N + ak(N-2)} \tag{1.1}$$

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 and 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.

1.3.5 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 et al. (2000) relax this assumption to study the effects of spatial population structure on the results of the UG. They analyse 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 neighbours). 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 neighbourhood 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 §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 neighbours 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), P235). Note that Alexander (2007) uses a mini UG contrarily to the models presented above, which could explain the discrepancy in the results.

Iranzo et al. (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, §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 neighbour's payoff through the offer that they make in the only interaction accepted between the pair. Hence, to avoid being at their neighbour'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.

1.3.6 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 (2002, 2001) 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 §), 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.

1.4 Terminological and theoretical problems

1.4.1 Terminological problems

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 et al., 2011, 2007). 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 et al. (2011, 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.

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 et al. (2014) report fairness for offers of 0.35, while Ichinose and Sayama (2014) describe offers as low as 0.25 as fair (see their Figure 1, p. 2, §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.

1.4.2 Theoretical concerns

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 1.3.2 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 SM1 section 4.1 (SM, 2015)). The results are presented in Figure 1 With $1 - p \geq q$, we replicate Nowak et al.'s results (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.

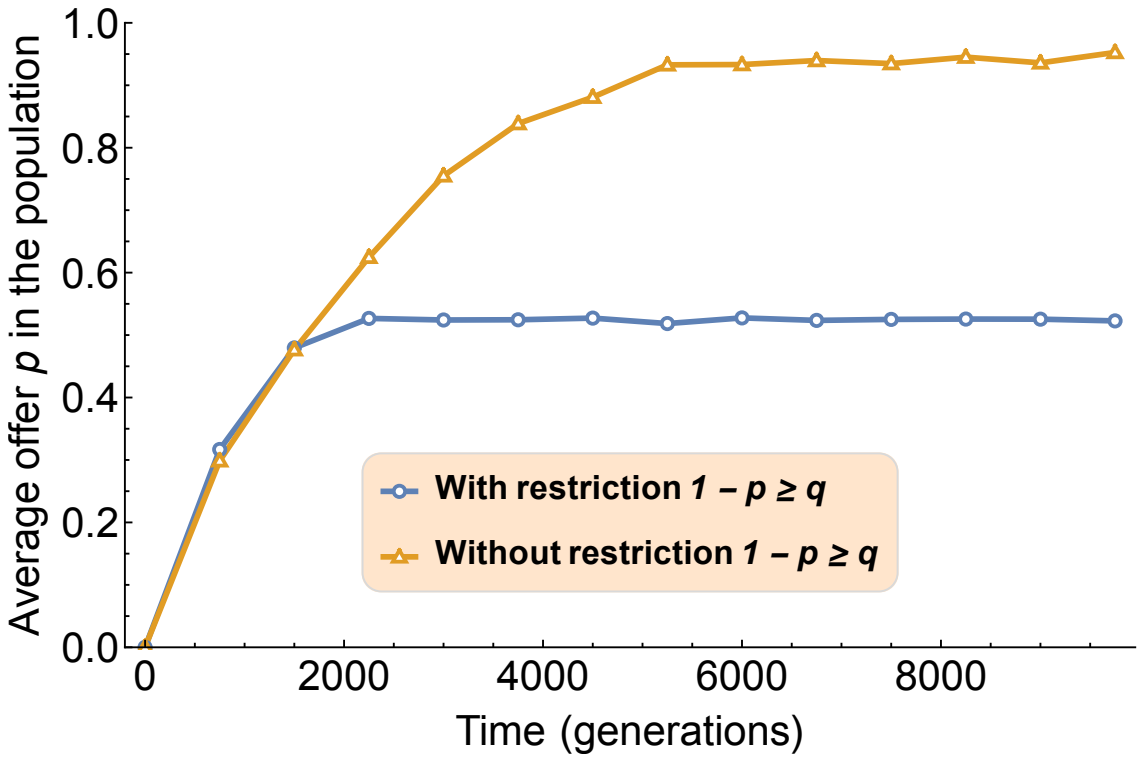

Figure 1: Evolution of the average ultimatum game offer in the model of Nowak et al. (2000), with and without the restriction $1 - p \geq q$. Only this restriction maintains offers around their fair value of 0.5. Each curve is an average over 20 simulation runs.

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 modelling 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, §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 can not evolve independently? Hence, we suggest 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.

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 and Samuelson (1994); Gale et al. (1995); Harms (1997), and see Skyrms (1996) for a discussion of discretizing evolutionary games to different degrees).

Empathy modelling

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 et al. (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 $p = q$ doesn't 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 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.

The importance of initial conditions

A few papers have reported the evolution of fairness to be dependent on the initial conditions of the model (Roth and Erev, 1993; Chiang, 2008), 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.

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.

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, §1), Wang et al. (2014) report the evolution of fairness "even in [an] information-deficiency situation" (p. 5 §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, §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 1.5.4) or to start comparing the models (see section 1.4.2). 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 (Leary and Kowalski, 1990; Haley and Fessler, 2005; Bateson et al., 2006), 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 1.5.4).

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 et al. (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 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.

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 1.3 (Nowak et al., 2000; Page et al., 2000; Page and Nowak, 2002; 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 favoured 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 1.3 using agent-based simulations: Nowak et al. (2000); Page et al. (2000); Page and Nowak (2002); André and Baumard (2011a); Rand et al. (2013). We will not analyse these simulations here, as a proper analysis would require an article of its own. We only make them available to the community, hoping some readers will find them useful. The complete models are available online at github.com, figshare.com, and the first author's personal website. 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 SM1 section 4 (SM, 2015). We hope 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 1.5.

1.5 Links between the theory and the experimental data

1.5.1 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 2 provides an overview of the timescales studied by each author, but many authors, with a few notable exceptions (Binmore and 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 SM1 section 1 (SM, 2015) 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 or 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 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 et al., 2008; Schmidt and Sommerville, 2011; Geraci and Surian, 2011; Warneken et al., 2011; Sloane et al., 2012), its neurological basis (Sanfey et al., 2003; Knoch et al., 2006; Tabibnia et al., 2008), its similarities with other species (Bräuer and Hanus, 2012; Warneken and Tomasello, 2009; Brosnan and de Waal, 2014), and, to a lesser extent, its universality (Marshall et al., 1999; Henrich, 2004). 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.

1.5.2 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, as 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 (Maynard Smith and Parker, 1976; Hammerstein, 1981). 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 frame-

work 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.

1.5.3 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 (Vehrencamp, 1983; Johnstone, 2000). 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 et al., 1991; Noë and Hammerstein, 1994). Models of asymmetric contests deal with the division of a resource when individuals differ in terms of competitive power (Maynard Smith and Parker, 1976; Hammerstein and Parker, 1982). Spite and spatial structure are two other mechanisms that have been widely investigated outside a human context (Gardner and West, 2004; Lehmann et al., 2006).

There is every reason for these models to be a great source of inspiration for human-related modelling, 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 modelling 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 2). 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.

1.5.4 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 in 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 1.3 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.

1.5.5 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 and Gächter, 2002; Gintis et al., 2003). Other scholars (Baumard and Sperber, 2010; Ceci et al., 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 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 can not 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 1.5.2 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.

1.6 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.

Chapter 2

The partner choice theory

"Well I wanted to maximize my own benefit, but I also wanted to throw the other person a bone. So I gave the other person 15% and kept the rest."

AN892ELDYF4GH

The theory of the evolution of fairness by partner choice has been presented in details in the two following publications:

- Baumard, N. *The Origins of Fairness: How Evolution Explains Our Moral Nature*, Oxford University Press (in press).
- Baumard, N., André, J. & Sperber, D. A mutualistic approach to morality: The evolution of fairness by partner choice. Behavioral and Brain Sciences. 6, 59–122 (2013).

I will only summarize the general argument and outline the important points for my thesis here, so I strongly recommend the interested reader to refer to the references above for a more detailed account, especially if my own account does not sound persuasive. In any case, this chapter can not be considered an original work of research.

2.1 Why an evolved sense of fairness?

Biologists usually have no problems to consider a biological origin to the sense of fairness, probably because the idea has been made familiar by the related work on

the evolution of cooperation. Other scholars can be more sceptical, because fairness seems restricted to the human species (see Chapter 8), and is often associated with religion, state institutions, or other cultural artefacts. I thus think useful to present the arguments which, if not peremptory, bring some support to the idea of a biological origin to the sense of fairness.

The general argument is that fairness seems to be, like other biologically-evolved modules, innate, universal, autonomous, and specific (Baumard, 2015).

Innate because fairness develops early in children (early enough that we can doubt that education or culture is entirely responsible for it). At 12 month old, children already react to unequal distributions (Geraci and Surian, 2011; Schmidt and Sommerville, 2011; Sloane et al., 2012). More sophisticated judgements such as proportionality between output and input appear later but still relatively early: 3 years-old can take merit into account (Baumard et al., 2012). Children also reject conventional rules when they go against fairness principles, which suggests they can make the difference between arbitrary cultural rules and fairness rules (Turiel, 2002).

Universal, because the anthropological data we have suggests that fair distributions can be found in many small-scale societies. If the sense of fairness is an adaptation to our ancient environment, it should be present in all current human populations, including hunter-gatherers who do not have the same religions or institutions we have in Western societies. Field observations show that primary distribution of game after a collective hunt is often shared according to the merit of each hunter (Gurven, 2004; Kaplan and Gurven, 2005; Alvard and Nolin, 2002). Harvested food such as roots or berries is also shared according to the effort (Gurven, 2004), and the developmental trajectory of fairness evoked in the previous paragraph has also been found cross-culturally (Liénard et al., 2013; Chevallier et al., 2015; Schäfer et al., 2015).

Autonomous, because fairness judgements are produced very quickly and automatically and are not based on conscious reasoning (see section 4.5).

Specific, because fairness can not be entirely reduced to other mental activities or mental states such as empathy, parental care, disgust, etc: fair judgements have their own logic that is different from the logic of other mental abilities (see sections 9.1 and 9.2 for a longer discussion of this point).

These arguments support the thesis of a biological origin to human fairness, but they are not enough to make it an adaptation. If it is already not easy to recognize an adaptation in evolutionary biology (Ridley, 2004), the task is even harder

in evolutionary psychology where the anatomy of the adaptation (and its possible engineering design) is not observable. Measurements of differential reproductive success are equally difficult. I see mostly two arguments that allow us to favour the adaptationist view. First, fair behaviors are costly (at least in the short-term). Hence, if they are so widespread in humans, there should be good reasons for why they have been kept over evolutionary times. Second, theoretical models such as the ones present in this thesis (Chapter 3 and 4) show that fair behaviors constitute the fitness-maximising behaviors in similar environments to the ones humans evolved in. In other words, the fined grained properties of fair behaviors measured empirically correspond to what we would predict to be the properties of an adaptation that would have evolved to solve the problem of dividing the costs and benefits of cooperation.

2.2 Why partner choice as an evolutionary mechanism?

Imagine you want to start studying the evolution of fairness, but you have no knowledge of the literature on the evolution of cooperation or prosocial behaviors in general. Is there a good reason why you would want to pick partner choice as your first candidate to study the evolution of fairness? I will argue that there is one indeed.

Philosophers have long identified the logic of "impartiality" that characterizes fairness. Contractualist philosophers use the metaphor of a contract to describe fairness: when they behave in a fair way, people behave as if they had previously agreed over a contract that would be mutually-beneficial for all parties. It turns out that partner choice can "naturally" explain this logic of impartiality, through the marginal value theorem (Charnov, 1976). The marginal value theorem tries to answer the following question: when individuals can forage and gain energy from different patches, how much time should they devote to each patch before moving on to another one? The answer is intuitive to grasp: individuals should stop foraging on a patch when the energy intake drops below the energy intake they could get elsewhere (taking into account the travel costs). In other words, the optimal foraging strategy when individuals can *choose* between patches is a strategy in which each unit of time invested in a patch brings at least the same returns than a unit of time invested in another patch. The analogy with social interactions becomes almost transparent: when people can choose between different social partners, they should not accept to interact with people that will give them less than what they could obtain elsewhere. Only mutually-beneficial interactions can be accepted in a partner

choice environment.

I would like to illustrate the strong connection between the philosophers' description of fairness and our work with a quick example. In an extension to Rawls' theory of justice, Dworkin suggests that to be called fair, any distribution should pass the "envy test". As Kymlicka (2002) puts it:

> *Dworkin asks us to imagine that all of society's resources are up for sale in an auction, to which everyone is a participant. Everyone starts with an equal amount of purchasing power-100 clamshells, in his example- and people use their clamshell to bid for those resources that best suit their plan of life. [...] [At the end of the auction,] everyone will be happy with the result, in the sense that they do not prefer anyone else's bundle of good to their own. If they did prefer a different bundle, they could have bid for it, rather than the goods they bid for.*

Once again, the analogy is striking. In real life, the evolutionary problem has not been to allocate clamshells to different bids but to allocate units of time to different social partners. But the principle that an equal endowment of clamshells/time leads to fair distributions when people can freely choose where to invest their resources is the same. That this idea of investment into different activities, that will be at the basis of many of our models (see Chapters 3 and 4), is already present in the philosophers' work is striking to me. Hence, from a purely theoretical point of view, it makes a lot of sense to consider partner choice as an explanation for the impartial contract-like aspect of human fairness. Finally, note that while contractualist philosophers have often been criticized for postulating that self-interest is at the basis of fairness (because individuals bid for their self-interest), which does not match the disinterested empirical nature of fairness, an evolutionary perspective easily solves this paradox: evolutionary approaches can accommodate self-interested motives at the ultimate level and a genuine care for others' interest at the proximate level (Baumard et al., 2013).

2.3 The importance of outside options

We can now reconstruct a plausible evolutionary history for fairness. Because of the marginal value theorem, natural selection favors individuals who share the costs and benefits of cooperation impartially, i.e. fair individuals. Fair individuals reap two benefits: being preferentially chosen as social partners but also not giving up too many resources. This last point is crucial: the "impartial" character of fairness could

not be explained if the only benefit of being fair was to attract social partners. In this case, we would predict fairness to be synonym of over-generosity (see Chapter 6 and 10.2 for a thorough discussion of this point). In other words, both the necessity to be chosen (working against selfish impulses) and the possibility to choose (working in the direction of selfishness) matter to explain the logic of fairness.

This distinction brings us to the important topic of outside options. Refusing to cooperate with an unfair partner is interesting as long as one has other opportunities to interact with more fair-minded individuals - as long as one has good outside options. In other words, partner choice in itself is not enough to explain why interactions should be fair: the fair or unfair status will ultimately depend on the outside options available to each individual. In previous models (André and Baumard, 2011a), fairness understood as equal divisions of resources evolves because all individuals have the same average outside options. In real life, we would expect that different individuals could be endowed with different outside options. Some individuals, because they are physically stronger for instance, could have on average better outside options than others. Some individuals could also have better skills and enjoy privileged outside options because of these skills. What are the best ways to reward social partners in this situation? What distributions are favored by natural selection? And do they match distributions that people empirically find fair? Answering these questions is the goal of the next Part.

Part III

Evaluating the explanatory power of partner choice

Chapter 3

Partner choice explains why it is unfair to exploit weaker people (Paper 2)

"I consider myself a fair person. I also would want someone to do the same for me, and I think that my actions should reflect my inner values."

ASLOM792N3M

3.1 Objectives and summary

For centuries, philosophers have discussed and denounced the fallacy of the "law of the strongest." In the first chapter of Plato's *Republic*, for instance, Thrasymachus claims that "justice is nothing else than the interest of the stronger," which Socrates then disputes. Many years later, in his foundational work on political rights, Rousseau noted that he could not see "how morality could result from the effects of physical power" (Rousseau, 1762). In the last decades, scientific research has accumulated evidence of humans' strong preference for egalitarian outcomes independently of the power relationship between individuals (Fehr and Schmidt, 1999; Fehr and Fischbacher, 2003; Camerer, 2003; Boehm, 1993, 1997; Dawes et al., 2007; Tricomi et al., 2010). This phenomenon is so universal that anthropologists have coined the term "egalitarian syndrome" to describe the prevalence of such preferences for equality in small-scale societies (Boehm, 1993, 1997; Cashdan, 1980). More generally, we take modern mass movements such as the anti-slavery, anti-discrimination, civil rights, and fair trade movements as expressions of the same urge to care for and defend the interests of the weak[1].

[1]Which also means that humans are capable of slavery, discrimination and unfair trade. I do not want to appear too naive about human behavior at this point, but I will not discuss unfairness

In our day-to-day life too, I suspect there are many situations in which we could take advantage of our strength but we do not do it; in fact, we do not even notice that we could do so (like not paying for our newspaper when the newsvendor is old and weak). On the contrary, when unfairness towards the weak does happen, it constitutes some of the most revolting situations we can be confronted to (like a sickly kid getting his snack stolen by the elder in the playground, or the lying old tramp getting beaten by Alex's gang in Stanley Kubrick's *A Clockwork Orange*).

Our previous models of the evolution of fairness by partner choice (André and Baumard, 2011b,a) were unable to account for this urge to defend the rights of even the weakest. André and Baumard (2011a)'s result relies heavily on the assumption that the dominant or subordinate status of an individual is randomly decided in each new interaction. In other words, in this scenario, the equal division of resources is the result of equal outside options. But the biological plausibility of equal outside options is highly debatable. For instance, in any scenario in which dominance is not random but linked to some intrinsic property of the individual, such as physical strength, individuals who are dominant in an interaction will also be more likely to be dominant in other interactions. This means that two individuals engaged in an interaction will not have the same outside options. Hence, it is legitimate to wonder whether the evolution of equal divisions will hold after the introduction of asymmetries of strength among individuals.

This section thus puts the partner choice theory to the test to see if it can explain the evolution of equal divisions, even after the introduction of asymmetries of strength among individuals. The answer is positive. In brief, the model shows that at the evolutionary equilibrium, strong individuals refrain from taking advantage of their strength to extort benefits from the weak, as long as weak individuals can abstain from interactions with bullies and instead cooperate with each other.

The following results have been published in:

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

much in my thesis (see nonetheless sections 4.5 and 10.2). I refer to Baumard (2015) for such a discussion. For now, I think it is enough to emphasize that I do not claim that humans never take advantage of their strength, but rather that refusing to take advantage of one's strength is a situation that happens often enough that it deserves to be studied and explained.

Evolution of equal division among unequal partners

Abstract: One of the hallmarks of human fairness is its insensitivity to power: while strong individuals are often in a position to coerce weak individuals, fairness requires them to share the benefits of cooperation equally. The existence of such egalitarianism is poorly explained by current evolutionary models. We present a model based on cooperation and partner choice that can account for the emergence of a psychological disposition toward fairness, whatever the balance of power between the cooperative partners. We model the evolution of the division of a benefit in an interaction similar to an ultimatum game, in a population made up of individuals of variable strength. The model shows that strong individuals will not receive any advantage from their strength, instead having to share the benefits of cooperation equally with weak individuals at the evolutionary equilibrium, a result that is robust to variations in population size and the proportion of weak individuals. We discuss how this model suggests an explanation for why egalitarian behaviors towards everyone, including the weak, should be more likely to evolve in humans than in any other species.

3.2 Introduction

For centuries, philosophers have discussed and denounced the fallacy of the "law of the strongest." In the first chapter of Plato's *Republic*, for instance, Thrasymachus claims that "justice is nothing else than the interest of the stronger," which Socrates then disputes. Many years later, in his foundational work on political rights, Rousseau noted that he could not see "how morality could result from the effects of physical power" (Rousseau, 1762). In the last decades, scientific research has accumulated evidence of humans' strong preference for egalitarian outcomes independently of the power relationship between individuals (Fehr and Schmidt, 1999; Fehr and Fischbacher, 2003; Camerer, 2003; Boehm, 1993, 1997; Dawes et al., 2007; Tricomi et al., 2010). In economic games, it has been shown that in collaboratively meaningful contexts, people favor equal divisions when contributions are equal (Dawes et al., 2007; Cappelen et al., 2007; Frohlich et al., 2004). A similar phenomenon is detectable early on in children, with three-year-olds dividing rewards equally after collaboration (Warneken et al., 2011). Finally, cross-cultural studies have shown that such behavior can be observed in many different cultural contexts, from small-scale societies to large industrial societies. In fact, this phenomenon is so universal that anthropologists have coined the term "egalitarian syndrome" to describe the prevalence of such preferences for equality in small-scale societies (Boehm, 1993, 1997; Cashdan, 1980). More generally, modern mass movements such as the anti-slavery, anti-discrimination, civil rights, and fair trade movements are all expressions of the same urge to care for and defend the interests of the weak.

Evolutionarily speaking, these observations raise the question: why should natural selection favor equal divisions of benefits, independently of the power struggle between the protagonists? Or, said differently, under what conditions is it adaptive for stronger or dominant individuals to leave half of the resource to their partners, when they could keep everything for themselves?

A useful paradigm for studying this question is the ultimatum game (UG). In this game, two individuals bargain over the division of a benefit, with one individual (the "proposer") making an offer to the other individual (the "responder"). If the responder accepts the offer, it is implemented; otherwise, none of the individuals receives any benefit. The very structure of this game implies an asymmetry of bargaining power between the two players. On the one hand, whatever offer a responder's partner makes, accepting it brings a greater gain than rejecting it. Therefore, in all cases, natural selection favors indiscriminate responding, with responders taking whatever benefits are made available to them. On the other hand, and as a result, selection favors stingy proposers, offering the minimal possible amount. The

division of benefits at the evolutionary equilibrium is thus maximally "unfair": the empowered individual (the proposer) keeps virtually all the benefits. The UG is therefore a conservative paradigm for studying the evolution of fairness.

In asymmetric interactions of this sort, in which a dominant individual can unilaterally impose a division of resources on another, both bargaining theory (in economics) and reproductive skew theory (in behavioral ecology) show that the dominated individual's outside options limit the level of inequality that the dominant can impose (Vehrencamp, 1983; Johnstone, 2000; Muthoo, 1999). However, these models treat the value of outside options as *exogenous* parameters, which are fixed a priori. Hence, whereas they can account for the fact that dominant individuals leave "something" to subordinates, they cannot explain quantitatively why *equal* divisions should precisely be favored by natural selection. André and Baumard (2011a) went a step further by showing that if outside options consist in the possibility of entering into another identical interaction with a new partner, and if the dominant or dominated status of an individual is randomly decided in each interaction, then each individual is certain to receive an expected payoff equal to half of the resource in the next interaction she will enter (see also André and Baumard (2011b) for a different model leading to the same consequence). This outside option thus forces dominant individuals to always share benefits in two equal parts, a mechanism that was already suggested verbally by Vehrencamp (1983) and that is also conceptually similar to the infinite-horizon, alternating-offers bargaining game of Rubinstein (1982).

André and Baumard (2011a)'s result relies heavily on the assumption that the dominant or subordinate status of an individual is randomly decided in each new interaction. In other words, in this scenario, the equal division of resources is the result of equal outside options. But the biological plausibility of equal outside options is highly debatable. For instance, in any scenario in which dominance is not random but linked to some intrinsic property of the individual, such as physical strength, individuals who are dominant in an interaction will also be more likely to be dominant in other interactions. This means that two individuals engaged in an interaction will not have the same outside options. Hence, it is legitimate to wonder whether the evolution of equal divisions will hold after the introduction of asymmetries of strength among individuals. If not, the observation that humans share equally even with weaker individuals, a central characteristic of fairness, would require another explanation.

Note that other scholars have proposed alternative explanations for the evolution of human fairness not based on the possibility of changing partners (Nowak et al.,

2000; Gale et al., 1995; Rand et al., 2013). We will present these explanations in the discussion and compare them with our own approach, but at this stage it is important to note that none of these alternative approaches takes into account possible asymmetries of strength between bargaining individuals. To our knowledge, this paper is the first theoretical study concerned with the evolution of human fairness that explicitly considers systematic asymmetries of strength. In the discussion, we will also highlight the limits of our model and its relationship with the non-human literature on biological markets (e.g., Noë et al., 1991; Noë and Hammerstein, 1995; Noë et al., 2001).

Here, we present both an analytical model and the results of individual-based simulations on the evolution of the division of a benefit in an ultimatum game-like interaction, in a population where individuals can change social partner. Individuals are assumed to be characterized by an intrinsic "strength" that affects their probability of playing the strategically dominant role of proposer in all UGs that they play. We investigated whether fair divisions evolve in such an environment, and in particular whether strong individuals refrain from taking advantage of their strength when they are paired with a weaker partner.

3.3 Methods

Individual-based simulations.

We consider a population of individuals who enter into a series of pairwise social interactions with random partners. All individuals begin their lives in a solitary state, and they then meet random social partners, among other solitary individuals, at a given constant rate β. When two individuals meet, one of them is given the role of proposer, with a probability that may depend on the relative strength of the two individuals (see below). The proposer offers a given division of benefits to the responder, who can then either accept or refuse. If the offer is accepted, the two individuals actually enter into the social interaction, which is assumed to take time. Hence, they leave the pool of available solitary individuals until the end of their interaction, which occurs at a constant "split" rate τ. On the other hand, if the offer is rejected, the two individuals immediately go back to the pool of solitary individuals without receiving any benefit. Note that although it is convenient to describe this interaction as an UG, it is not a real UG stricto sensu, as the responder always has the choice of refusing an offer and hoping to interact with someone else in the population, which is not the case in the UG.

We consider a Wright-Fisher population with non-overlapping generations. Each generation lasts for a constant number of time steps, at which point all individuals reproduce according to the amount of benefits they have accumulated throughout their life, and then die. Genetic recombination is allowed between each generation.

The cost of partner choice.

The cost of partner choice is implicit in the above model. It is a consequence of the time it takes to find a new partner after the rejection of an offer. Hence, the cost and benefit of being choosy are not controlled by explicit parameters, but by two parameters that characterize the "fluidity" of the social market: the "encounter rate" β, and the "split rate" τ. When $\frac{\beta}{\tau}$ is large, interactions last a long time (low split rate τ) but finding a novel partner is fast (high encounter rate β), and individuals thus should be picky about which offers they accept. On the contrary, when $\frac{\beta}{\tau}$ is low, interactions are brief but finding a novel partner takes time, and individuals should thus accept almost any offer.

Strength.

We assume individuals are characterized by an intrinsic quantitative property $\sigma \in [0, 1]$ representing their "strength", which affects their probability of playing the advantageous role of proposer in the UG. This intrinsic property is constant across the entire life of an individual but is not heritable: i.e., at birth, each individual is randomly attributed a given level of "strength" that is independent of his parent's, according to a random distribution (see below). In an interaction between two individuals, we assume that the strongest of the pair has a given constant probability $\frac{1}{2} * (1 + \phi)$ of playing the role of proposer, where $\phi \in [0, 1]$ is a constant parameter, independent of the quantitative difference between the partners' strengths. When $\phi = 1$, strength controls the attribution of roles deterministically: the stronger partner always plays the role of proposer. When $\phi = 0$, strength has no effect on the assignment of roles. We also assume that, when two individuals of *exactly* equal strength are paired together, they have an equal chance of playing the role of proposer.

Regarding the distribution of individual strength at birth, for the sake of simplicity in our analytical approach and in most of our simulations, we assume that there are only two strengths, and thus only two types of individuals ("strong" and "weak"). In this case, we will call any given individual's probability of being randomly designated as "weak" at birth x. In other versions of our simulations, we assume instead that the strength of an individual at birth is sampled from a uniform distribution

between 0 and 1. In this case, individuals never interact with a partner of the exact same strength.

The social strategy

To play a UG, each individual must be characterized by two different behavioral variables: the *offer* they make when they play the role of proposer, and their *request* as a responder, i.e., the minimum offer they are ready to accept from their partner. The aim here is to consider the possibility of individuals detecting their partner's strength and adapting their behavior accordingly.

With only two levels of strength in the population, we assume that individuals are characterized by eight genetic variables: four p_{ij} and four q_{ij} variables, with i and $j \in \{s, w\}$ denoting an individual's strength (s for "strong", w for "weak"). p_{ij} is the offer made by a proposer of strength i in an interaction with a responder of strength j. q_{ij} is the minimum offer that a responder of strength i is ready to accept in an interaction with a proposer of strength j. For example, a strong individual who is the proposer in an interaction with a weak individual will propose p_{sw} benefits to the responder. The weak individual will then compare the value of p_{sw} to his own q_{ws}, and if $p_{sw} \geq q_{ws}$, the offer will be accepted.

With a continuum of strength in the population, we assume the offer is controlled by three underlying genetic traits: a constant γ, a degree of linear dependence on the individual's own strength ρ_p, and a degree of linear dependence on the partner's strength ρ_r. The offer is given by:

$$Offer = \gamma + \rho_p * \sigma_p + \rho_r * \sigma_r$$

with σ_p being the strength of the proposer, σ_r being the strength of the responder, $\gamma \in [0, 1]$ and $\rho_p, \rho_r \in [0, 1]$. Correspondingly, with a continuum of strength the responder's request is genetically encoded by three loci μ, λ_p and λ_r, and given by the expression:

$$Request = \mu + \lambda_p * \sigma_p + \lambda_r * \sigma_r$$

with σ_p being the strength of the proposer, σ_r being the strength of the responder, $\mu \in [0, 1]$ and $\lambda_p, \lambda_r \in [0, 1]$.

Note that although the system of offers and requests is a convenient way to model these interactions, it can be interpreted biologically in a different and probably more realistic way. The existence of offers does not necessarily imply an underlying contract: offers can also mean that responders have some information on the proposer's

usual behaviors, for example through a reputation built up in the course of past interactions with other individuals. Therefore, when an individual is characterized by an offer p, we can also interpret this as this individual having the public "reputation" of offering p.

Analytical model.

The analytical model incorporates all of the features of the simulations presented above, but with one simplification: we assume that the total number of interactions accepted per unit of time is the same for each individual. With this assumption, rejecting an opportunity to cooperate does not compromise the chances of cooperating later, but on the contrary grants new opportunities. This situation is analogous to the condition where $\frac{\beta}{\tau}$ tends towards infinity in the simulations: social opportunities are plentiful at the scale of the length of interactions. When individuals reject an interaction, however, they are forced to postpone their social interaction to a later encounter. We assume that this entails an explicit cost expressed as a discounting factor δ ($0 \leq \delta < 1$). If we call the average payoff of an individual of strength i G_i, then δG_i will be the average expected payoff in the next interaction after rejecting an offer. When δ equals 1, refusing an interaction carries no cost; when δ equals 0, refusing an offer will result in zero payoff from the next interaction. In practice, we will neglect the case where δ equals 1, as it leads to artefactual results (see SM2 section 1.2 (SM, 2015)). The analytical model is fully explained and solved in SM2 section 1.2 (SM, 2015).

The question we want to answer is the following: how will offers and requests evolve in such an environment, where individuals of different strengths coexist and share resources? Unless otherwise specified, the following results are concerned with the case where there are only two strengths coexisting in the population.

3.4 Results

When the population is made up of equal numbers of strong and weak individuals ($x = \frac{1}{2}$), if partner choice is not costly, the difference in strength between strong and weak individuals has little impact on the offers that strong individuals make to weak individuals (Fig. 2). If partner choice is not costly, starting from a stingy population of strong individuals offering nothing to weak individuals, offers progressively raise in the population up to the point where the strong offer close to half of the resource to their weak partners (Fig. 2 circle markers). In fact, strong individuals offer weak individuals as much benefit as they offer to other strong individuals when partner

choice is not costly (SM2 Fig.1, SM 2015). On the contrary, when partner choice is highly costly (Fig. 2 diamond markers), strong individuals make very low offers to weak individuals. This result holds even if there is a continuum of strengths in the population (not just two; SM2 Fig. 2 (SM, 2015)). As long as partner choice is not too costly, at the evolutionary equilibrium individuals who are paired with a stronger individual and playing the role of responder will still receive half of the total resource to be shared.

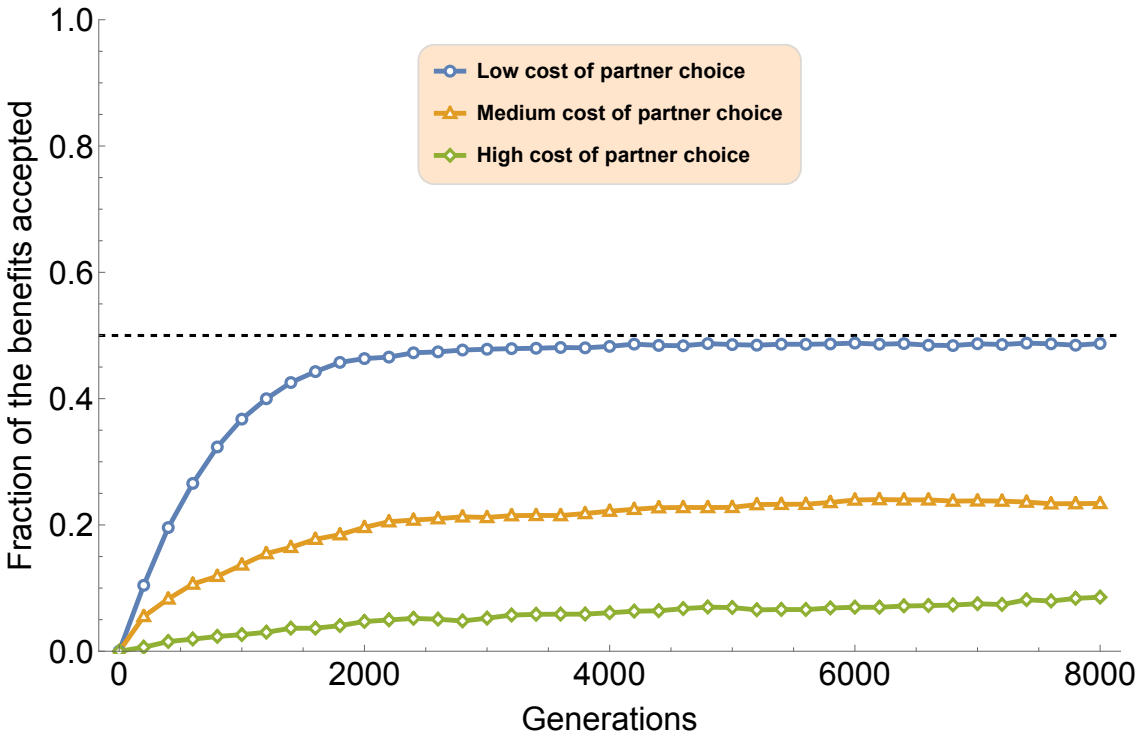

Figure 2: Average offer accepted by a weak individual paired with a strong individual across generations (simulation results). Average over 30 simulations. The dashed line corresponds to the (theoretical) perfectly equal division (50 %). Parameter values: $\phi = 1, x = \frac{1}{2}$. Circle markers: partner choice has almost no cost ($\frac{\beta}{\tau} = 100$). Triangle markers: partner choice has a medium cost ($\frac{\beta}{\tau} = 1$). Diamond markers: partner choice is highly costly ($\frac{\beta}{\tau} = 0.01$). When partner choice is not costly, weak individuals only accept offers that are close to 50% at the equilibrium. Other parameters used for these simulations can be found in SM2 section 1.6 (SM, 2015).

In the previous results, we arbitrarily set the proportion of weak individuals (x) at 0.5. It is plausible to think that this parameter will influence the division of benefits, since it impacts the social opportunities of weak individuals. To determine if equal divisions can still evolve when there is a low proportion of weak individuals in the population, we ran simulations for different values of x. The results show that this parameter in fact has a very limited fimpact: divisions of resources between strong

and weak individuals continue to be equal when the percentage of weak individuals is as low as 5% (Fig. 3 left panel).

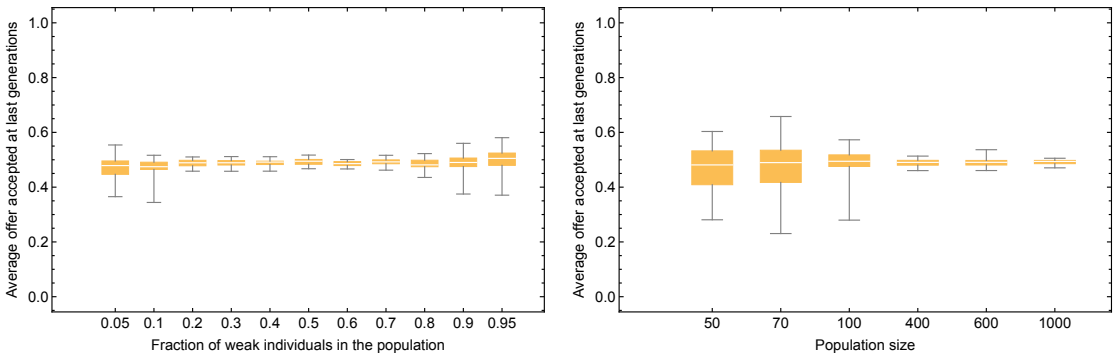

Figure 3: Robustness of the evolution of equal offers between strong and weak individuals (simulation results). Left panel: average offer accepted by a weak individual paired with a strong individual at the evolutionary equilibrium, for different fractions of weak individuals in the population. Right panel: average offer accepted by a weak individual paired with a strong individual at the evolutionary equilibrium, for different population sizes. Average over 30 simulations; $\frac{\beta}{\tau} = 100; \phi = 1$. In the left panel, $x = \frac{1}{2}$. Whiskers represent the min and max value obtained across 30 runs. Additional parameter values can be found in SM2 section 1.6 (SM, 2015).

Population size is another parameter that could affect the payoff to the weak: the smaller the population size, the smaller the total amount of social opportunities available to each individual. To test the effect of this parameter, we analyzed the payoff to weak individuals at the evolutionary equilibrium in populations of different sizes. Population size plays a role in determining the payoffs received by weak individuals, but quasi-equal divisions can be found in populations as small as 50 individuals (Fig. 3 right panel).

The results of the analytical model confirm the simulation results. When partner choice is not costly, at the evolutionary equilibrium strong individuals do not take advantage of their strength to offer unequal divisions to weak individuals (Fig. 4). Even in the case where weak individuals are *always* in the strategically dominated position of responder when paired with a strong individual ($\phi = 1$), they will receive half of the resource at the equilibrium as long as the cost of changing partners is not too high.

Analytical results also confirm that the frequency of weak individuals in the population has a small impact on divisions at the evolutionary equilibrium. Figure 5 shows the offer made by a strong individual to a weak individual at the evolutionary

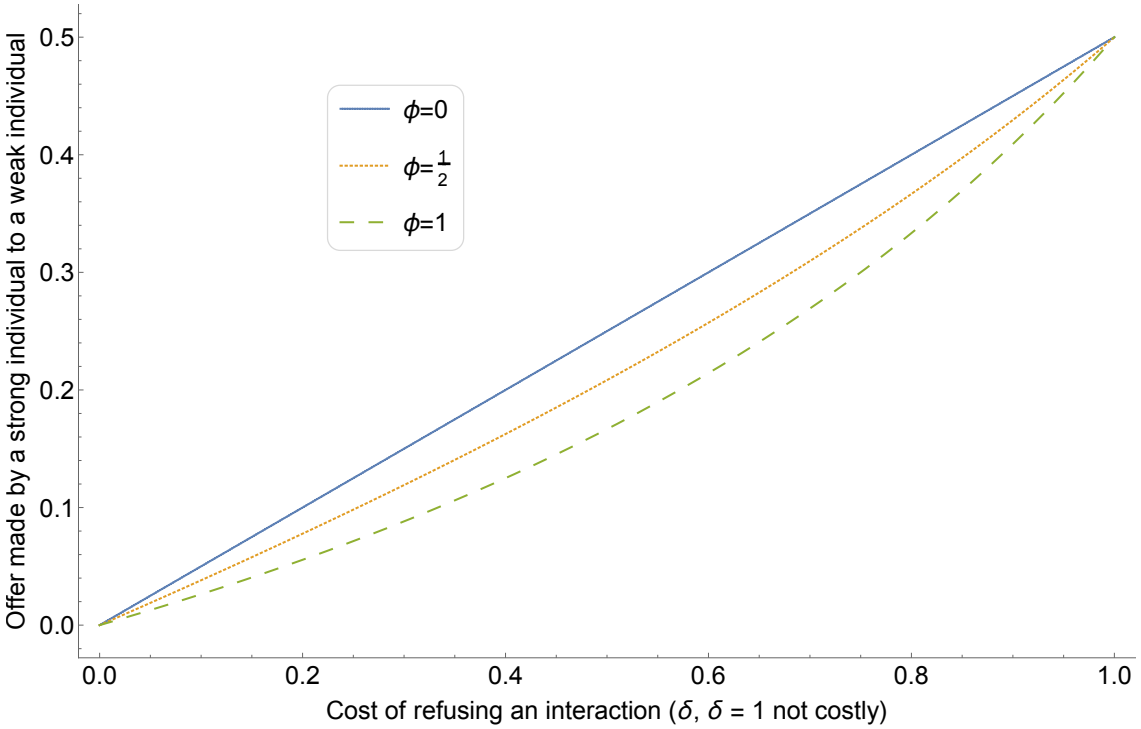

Figure 4: Average offer made by a strong individual to a weak individual at the evolutionary equilibrium, as a function of δ and for three values of ϕ (analytical results). Resource size is 1, $x = \frac{1}{2}$. The degree of dominance of strong individuals ϕ has a small impact on the offer they make. When partner choice is not at all costly ($\delta \to 1$), weak individuals receive half of the total resource.

equilibrium, when $\phi = 1$, for different values of x. As long as partner choice is not too costly, weak individuals will receive close to half of the resource to be shared, even when there are not many of them in the population ($x \to 0$). However, the higher the cost of changing partners, the more restrictive the parameter x becomes.

3.5 Discussion

In this article, we have shown that equal divisions can evolve in an interaction similar to the ultimatum game even when some individuals are stronger than others in the population, and thus have better average outside options than other individuals. Although they have a strategic bargaining advantage, strong individuals agree to give close to half of the benefits of interactions to weak individuals at the evolutionary equilibrium, a result that is robust to variations in population size and in the proportion of strong individuals in the population. To our knowledge, this is the first theoretical study on the evolution of human fairness that explicitly considers

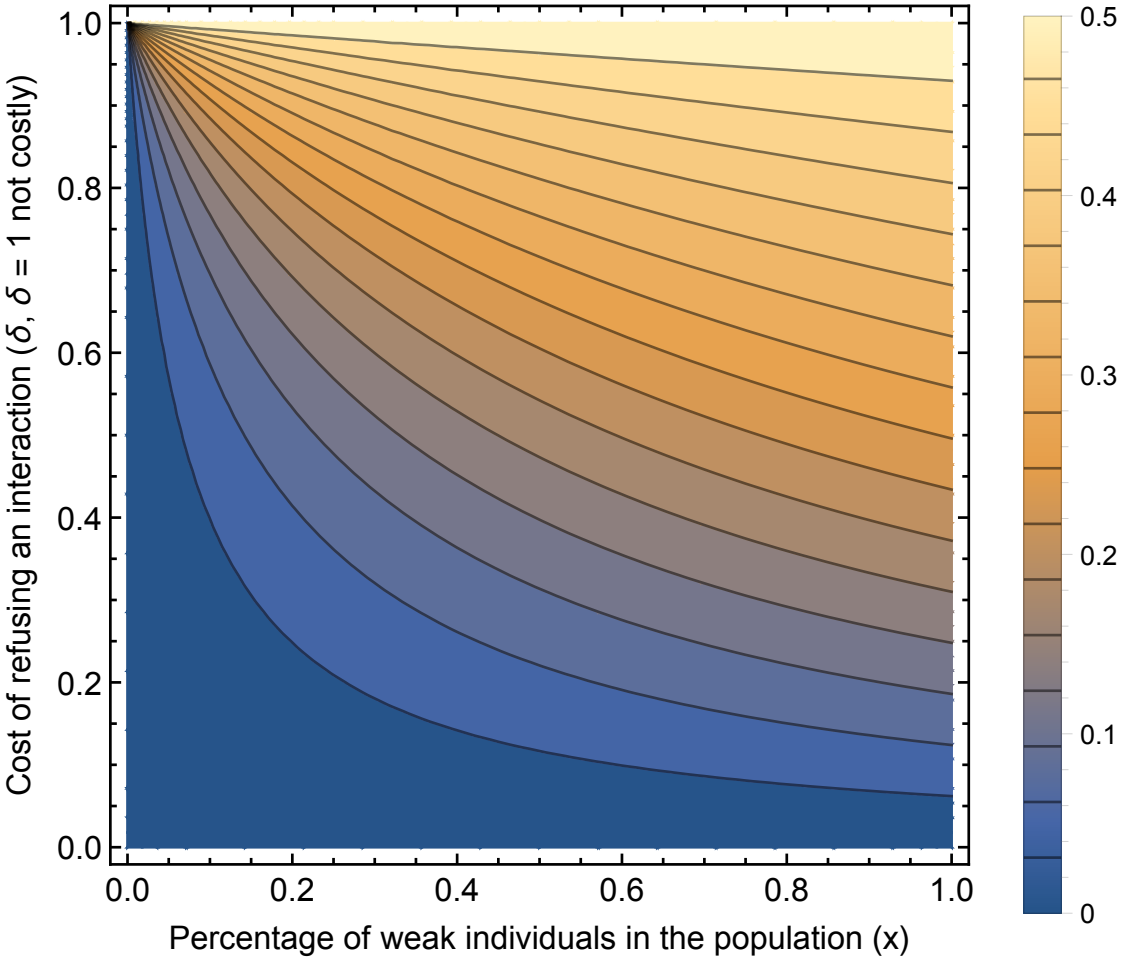

Figure 5: Average offer made by a strong individual to a weak individual at the evolutionary equilibrium as a function of δ and x (analytical results). Resource size is 1, $\phi = 1$. The percentage of weak individuals in the population has a significant effect on the weak individual's payoff, but this effect is limited when partner choice is not costly.

systematic asymmetries of strength, and shows that strength is not an evolutionarily relevant parameter to determine the division of benefits in an environment where partner choice is possible. In particular, we relax a crucial assumption used in previous modelling approaches (André and Baumard, 2011a) to show that equal divisions can still evolve when some individuals have better average social opportunities than others, a condition necessary for understanding the reach of partner choice-based fairness in humans.

These results shed light on an interesting question: why do strong individuals not more often take advantage of their strength to exploit weak individuals? The answer seems to be that the advantage of strength is only *local*, when what matters

when individuals can choose their social partners are *global* social opportunities. When individuals are embedded in a rich network of cooperative interactions and social opportunities, their bargaining power is determined not by local dominance relationships in each interaction, but by their outside options in the population as a whole. Thus, a weak individual's bargaining power is not affected by being locally dominated in a pairing with a strong individual, because of all the social opportunities that are available with other partners. Confronted with an unfair offer from a strong individual, a weak individual can easily refuse it and wait for an encounter with another weak individual. As a consequence, a strong individual who wants to interact with a weak individual will have to offer at least what the weak individual could gain elsewhere.

It is important to note that this result holds even when the relative proportion of weak individuals in the population is small. Variation in the number of weak individuals does indeed affect the average social opportunities of the weak. However, as long as a weak individual has at least one other weak partner in the population, this potential cooperative opportunity will constrain what strong individuals can offer. This result highlights the fact that the factor that most determines individuals' payoffs in an environment with varied social opportunities is not the *average* of the social opportunities that are available to them, but their *best* social opportunities. Weak individuals can "put forward" their best social opportunities, in which they can gain $\frac{1}{2}$ on average, when they are bargaining with strong individuals, which leads to the evolution of equal divisions even in strongly unbalanced populations.

Note that our results could also be interpreted the other way around: if partner choice is not sustained, the evolutionarily stable strategy will not be fair. If human beings have the ability to adapt (plastically) their level of fairness to cues indicating the efficiency of partner choice, then this could help to explain some of the inequalities we observe in pastoralist or agriculturalist societies, going from mild stratification to extreme cases of despotism and slavery (Kaplan et al., 2009; Summers, 2005), or inequalities in modern societies, for example in situations of monopoly (Kahneman et al., 1986a; Piketty and Saez, 2014). Although these links are purely speculative at this point, someone interested in the history of inequalities could make interesting predictions with our model.

Our model is related to existing models of the evolution of human fairness, but previous models have not taken into account differences of strength between individuals, and thus cannot explain the emergence of equal divisions when power is unequally distributed. A widely-cited model by Nowak et al. (2000) suffers from a restriction of the parameter space that undermines its main result: offers pro-

gressively increase because responders can build up a reputation for refusing low offers, but offers stabilize at 50% only because the authors arbitrarily assume that proposers cannot request more as a responder than what they offer as a proposer, following an unsubstantiated assertion that "individuals do not regard the role of proposer as inferior to the role of responder" (Nowak et al. (2000) footnote 14; see André and Baumard (2011b) for a lengthy discussion of this problem). Another line of models shows that introducing high variance in the responders' acceptance threshold can also lead to increased offers in the ultimatum game (Gale et al., 1995; Rand et al., 2013; Ichinose, 2012). Variance, depending on the model, can be due to a high mutation rate, weak selection, learning mistakes, difficulty assessing the strategies of others, etc., but the general mechanism is the same: when noise leads a sufficiently large proportion of responders to keep refusing low offers, proposers have no choice but to increase their offers. A recent work also suggests that "spiteful" strategies negatively assorted with a mix of other strategies could lead to a certain degree of fairness (Forber and Smead, 2014), even though the use of the word "spite" in an evolutionary context is highly debatable (West and Gardner, 2010). To our knowledge, the only existing model on the evolution of equality with differences of strength is the model developed by Gavrilets (2012). In this model, the need for individuals to not only maximize their own fecundity but also minimize the fecundity of others can lead to the evolution of helping behaviors directed towards weak individuals engaged in agonistic ("owner-bully") interactions with stronger individuals. This last mechanism based on intense inter-individual competition is an alternative to the one we suggest based on intense cooperation, and it would be interesting to see how they compare when it comes to explaining finer-grained properties of human fairness (Konow, 2000; Baumard et al., 2013).

Overall, the scarcity of models of the evolution of human fairness incorporating differences of strength makes our approach more closely related to models of reproductive skew. Skew theory aims to explain why in some species the benefits of reproduction are highly skewed in favor of dominant individuals, whereas in other species a more equal division occurs and subordinates reproduce as much as dominants (Johnstone, 2000). Some "tug-of-war" models of reproductive skew have been applied to human cooperation (Barker et al., 2012), but only transactional models emphasize the important role of subordinates' outside options in determining how reproduction is divided. They show that when subordinates have good outside options, dominants are forced to give them a large number of reproductive opportunities if they want to keep them in their group (Vehrencamp, 1983; Reeve et al., 1998). Theoretically, we depart from reproductive skew models in the way we model outside options: rather than arbitrarily fixing a certain value for them,

outside options in our model emerge from the dynamics of social interactions themselves, i.e., from encounters between individuals and the cost of changing partners. In this situation, if partner choice is not costly, the model shows that strong individuals not only give *something* to weak individuals, they give exactly half of the resource to be shared. Our model is also strongly inspired from biological markets models (Noë et al., 1991; Noë and Hammerstein, 1994), in which commodities are being exchanged and trading partners compete to be chosen by the other trading class. A general result of this line of models, reminiscent of our own results, is that supply and demand represented by the trading classes will determine the value of the exchanged commodity (Noë and Hammerstein, 1995; Noë et al., 2001; Johnstone and Bshary, 2008). An important difference with our model though is that we do not suppose an agent has to be firmly attached to a specific trading class. Whereas it makes sense when modeling mating markets or interspecific mutualisms to assign a fixed class to each agent (male or female, species A or species B, breeder or helper, etc.), here we investigated what happens when individuals can freely switch from one class to the other.

One important question concerns the biological interpretation of the "strength" postulated in our models. Because we chose to model strength abstractly, it can represent any feature that affects individuals' "resource-holding potential" as described classically in behavioral ecology (Arnott and Elwood, 2009): body size, body mass, development of weaponry, physiological state, etc. The central property of strength in our models, however, is that it brings individuals only a local bargaining power advantage, and not a global one. Fairness evolves in spite of asymmetries of strength if stronger individuals are more likely to dominate each interaction, but not if they can actively reduce the outside options of weaker individuals by preventing them from cooperating with each other. This definition of strength constitutes a clear limit to our model: if we assumed that stronger individuals had the possibility to coerce their partner into interacting, fairness could not evolve. However, reducing others' outside options requires a much higher investment than just taking away resources, as individuals need to monitor who else their partners are interacting with at any given moment. To do so, strong individuals would have to spend the majority of their time guarding others, to the detriment of the production of cooperative benefits. At the very least, we predict that, in situations in which such partner guarding is impossible, the benefits of cooperation should be divided equally, irrespective of the resource-holding potential of individuals. Another limit of our model is that we suppose strength is not heritable (i.e. we assume that investment into competitive strength cannot evolve by natural selection). There is no doubt it would be fruitful to relax this simplifying assumption in further models to see whether an interesting

pattern of coevolution between strength and fairness emerges.

While cooperation and partner choice are not restricted to humans, egalitarianism seems to be rare outside the human species. Chimpanzees, unlike children for instance, do not share benefits equally even when they had to collaborate to produce them (Melis et al., 2006), and rarely share food at all in natural settings (Tomasello et al., 2012). Both within and beyond the primate order, high-ranking males usually enjoy more resources than low-ranking males (Ellis, 1995). In the kingdom Animalia, contests over resources are most often won by the individual with the highest resource-holding power (Arnott and Elwood, 2009). This raises the question: why are humans so prone to respect the interests of the weak? Although our model does not allow us to answer this question with certainty, it offers at least two different hints. The first is that weak individuals in the human species may have a better choice of cooperative partners than those in other species. The second hint is that in humans, strength may have a far lesser role in the generation of benefits than it is in other species. Because of the nature of human cooperation and the variety of forms it can take, there are ways for two physically weaker individuals to produce benefits equivalent to one weak and one strong individual working together (Wiessner, 1996; Kaplan et al., 2009). In other words, strength ceases to be an important factor in determining the *division* of benefits because it ceases to be an important factor in determining the *production* of benefits. In other species, on the contrary, the involvement of a strong individual almost always brings extra benefits that cannot be produced in interactions involving only weak individuals (protection, access to females, territory, or food). As a consequence, distributions skewed in favour of strong individuals are much more frequently observed in non-human animals (Ellis, 1995; Grafen, 1987). Although at this point these two non-exclusive hypotheses are speculative and cannot be confirmed theoretically, our model suggests that both the quantity of human cooperation (meaning that individuals always have a rich overall set of social opportunities) and its nature (allowing even physically weaker and hence less competitive individuals to produce similar benefits) are human-specific selection pressures that could have led to the evolution of concern for the interests of the weak. Of course, human egalitarian behaviors have varied greatly across time and space, and our simple model based on genetic transmission alone cannot capture the full complexity of human behavior and cognition in relation to cooperation. It may nonetheless be able to explain the general and universal pattern of egalitarian behaviors observed in many human societies (Boehm, 1993, 1997; Cashdan, 1980).

Chapter 4

Partner choice explains why it is fair to reward according to contribution (Paper 3)

"I thought that I deserved a little more than the other person since the other person didn't have to do anything to get the points."

ANS892N2CD4L

4.1 Objectives and summary

If the necessity to defend the rights of the weak is one of the hallmarks of human fairness, the necessity to reward people according to their contribution might be another one. Aristotle for instance suggested an "equity formula" (Aristotle, 1999) for fair distributions, mathematical equivalent of "reward according to contribution", whereby the ratios between the outputs O and inputs I of two persons A and B are made equal: $\frac{O_A}{I_A} = \frac{O_B}{I_B}$. Today, it is well accepted in the behavioral sciences that people prefer income distributions with strong work-salary correlations, prefer to give more to individuals whose input is more valuable, and favor meritocratic distributions as a whole in both micro- and macro-justice contexts (Baumard et al., 2013). In economic games, for instance, participants consistently divide the product of cooperative interaction in proportion to each individual's talent, effort, and the resources invested in the interaction (Cappelen et al., 2010; Frohlich et al., 2004). "Reward according to contribution" is also at the center of equity theory, the behavioral psychology theory of fair and unfair distributions (Adams, 1963; Adams and Jacobsen, 1964). Meritocratic distributions have been observed across many societies (Marshall et al., 1999), including hunter-gatherer societies (Gurven, 2004; Alvard, 2004). For instance, Bailey (1991) (cited in Baumard et al. (2013)) describes

how primary distribution takes place after collaborative hunt in Efe pygmies: the hunter who shot the first arrow receives on average 36% of the prey, the owner of the dog who chased the prey 21%, the hunter who shot the second arrow 9%, etc[1]. Meritocratic distributions can also be detected early in human development (Kanngiesser et al., 2010; Baumard et al., 2012), with 3 years-old children recognizing that the fair thing to do is to give more to people who contributed more to a cooperative action.

Hence, as was the case in the previous chapter, if partner choice is to be a good candidate to explain the evolution of human fairness, it needs to be able to explain the pattern of proportionality between inputs and outputs that we observe in fairness judgments. The goal of this chapter is to test this prediction. From a theoretical point of view, this study is also interesting because as I mentioned in Chapter 1, almost all studies of the evolution of fairness study fairness-in-the-ultimatum-game, i.e. the evolution of equal divisions. Finding out which theories can explain more than equal divisions (which can be seen as a particular case of proportional divisions when contributions are the same) is thus a good way to decide between the theories.

Simple agent-based simulations, a game-theoretical model, and extended simulations with evolving neural networks provide converging support for the conclusion that when individuals can choose their cooperative partners, meritocratic distributions emerge as the best strategy.

The rest of this chapter comes from an article currently under review in *Evolution and Human Behavior*.

[1]It is also well described that secondary distribution (once returned to the camp) of game among hunter-gatherers does not seem to be based on merit, with good hunters agreeing to receive relatively smaller portions. I will discuss this kind of sharing in section 10.1.

On the evolutionary origins of equity

Equity, defined as reward according to contribution, is considered a central aspect of human fairness in both philosophical debates and scientific research. Despite large amounts of research on the evolutionary origins of fairness, the evolutionary rationale behind equity is still unknown. Here, we investigate how equity can be understood in the context of the cooperative environment in which humans evolved. We model a population of individuals who cooperate to produce and divide a resource, and choose their cooperative partners based on how they are willing to divide the resource. Agent-based simulations, an analytical model, and extended simulations using neural networks provide converging evidence that equity is the best evolutionary strategy in such an environment: individuals maximize their fitness by dividing benefits in proportion to their own and their partners' relative contribution. The need to be chosen as a cooperative partner thus creates a selection pressure strong enough to explain the evolution of preferences for equity. We discuss the limitations of our model, the discrepancies between its predictions and empirical data, and how interindividual and intercultural variability fit within this framework.

4.2 Introduction

For centuries, philosophers have emphasized the important role of proportionality in human fairness. In the fourth century BC, Aristotle suggested an "equity formula" for fair distributions (Aristotle, 1999), mathematical equivalent of "reward according to contribution," whereby the ratios between the outputs O and inputs I of two persons A and B are made equal: $\frac{O_A}{I_A} = \frac{O_B}{I_B}$. This formula also captures the concept of "merit," the idea that people who work harder deserve more benefits (Adams, 1963; Konow, 2003; Skitka, 2012).

Psychological research on distributive justice, and on equity theory in particular, has offered extensive empirical support for Aristotle's claim (Adams, 1963; Homans, 1958; Walster et al., 1973; Mellers, 1982). Equity theory aims to predict the situations in which people will find that they are treated unfairly. A robust finding is that receiving more or less than what one deserves leads to distress and attempts to restore equity by increasing or decreasing one's contribution (Adams, 1963; Adams and Jacobsen, 1964). People prefer income distributions with strong work-salary correlations, prefer to give more to individuals whose input is more valuable, and favor meritocratic distributions as a whole in both micro- and macro-justice contexts (Baumard et al., 2013).

More recently, experiments with economic games have shown that participants consistently divide the product of cooperative interactions in proportion to each individual's talent, effort, and the resources invested in the interaction (Cappelen et al., 2010; Frohlich et al., 2004). Meritocratic distributions have been observed across many societies (Marshall et al., 1999), including hunter-gatherer societies (Gurven, 2004; Alvard, 2002; Liénard et al., 2013; Schäfer et al., 2015), and can be detected very early in human development (Kanngiesser et al., 2010; Baumard et al., 2012), suggesting that equity could be a universal and innate pattern in human psychology.

Preferences for equitable outcomes present the same evolutionary problem as preferences for fair outcomes in general: at least in the short-term, those preferences are costly. Although people react more to inequitable situations when they are disadvantageous than when they are advantageous, people still feel uncomfortable in unjustified advantageous situations (Austin and Walster, 1974; Fehr and Schmidt, 1999). Experiments even show that people are ready to incur costs and decrease their own payoff in order to achieve more equitable distributions (Dawes et al., 2007). How can natural selection account for the evolution of such costly preferences ?

Until now, little attention has been given to this question. There have been many theoretical studies on the evolution of fairness (Nowak et al., 2000; Gale et al., 1995;

Page and Nowak, 2002; Barclay and Stoller, 2014; André and Baumard, 2011a; Debove et al., 2015a), but all of them are concerned with explaining the evolution of fairness in the ultimatum game, an economic game where the fair division happens to be a division into two equal halves (Güth et al., 1982; Camerer, 2003). However, equal divisions are just a special case of the more general category of equitable divisions: that is, divisions proportional to contributions. As emphasized by equity theory, unequal divisions can be judged fair when they respect the partners' investment, talents, commitment, etc. In brief, although many models can explain the evolution of preferences for *equal* divisions, none of them is able to explain the evolution of preferences for *proportional* divisions. Here we aim to understand whether natural selection can lead to such proportional divisions of resources (including the particular case of equal divisions), in a scenario where partners can make differing contributions to a cooperative undertaking.

Partner choice has had an important role in the evolution of cooperation, as evidenced by both theoretical (Aktipis, 2004; Nesse, 2007; Aktipis, 2011; McNamara et al., 2008; Barclay, 2011) and empirical studies (Barclay, 2004; Barclay and Willer, 2007; Sylwester and Roberts, 2013, and see Barclay, 2013 for a review in humans). When people are in competition to be chosen as cooperative partners, experiments show that they increase their level of cooperation because they have a direct interest in doing so (Barclay, 2004, 2006). Partner choice also has interesting consequences for the evolution of fairness. It leads to equal divisions in theoretical and empirical settings (André and Baumard, 2011a; Debove et al., 2015b,a), because when individuals can choose whom to cooperate with then they are better off refusing divisions that do not match what they can obtain elsewhere in their social environment. Nonetheless, none of these studies were concerned with the evolution of equitable (proportional) divisions.

To summarize, preferences for equity are robust and widespread in humans, but we currently lack an evolutionary explanation for their costly existence. Here, we aim to put the partner choice mechanism to the test to see if it can explain such preferences. We develop models in which individuals put effort into the production of a collective good, and differ with regard to both the amount of effort they are willing to put in and to the efficiency of their contribution to the production of the good. To determine the evolutionarily stable sharing strategy in this environment, we first analyzed an evolutionary model using agent-based simulations. We then developed an analytical model and performed simulations with evolving neural networks as more realistic decision-making devices. The results provide converging support for the conclusion that when individuals can choose their cooperative partners, equity

emerges as the best strategy, and the offers that maximize fitness are those that are proportional to the individual's relative contribution to the production of the good.

4.3 Methods

The model is an improved version of the partner choice model presented in Debove et al. (2015b), integrating individual differences in productivity and effort, and using neural networks as decision-making devices. Source code for the simulations is available online.

4.3.1 Simulations

Social life

Individuals randomly meet each other at a constant rate β in the population. When two individuals meet, one of them is randomly selected to decide how the resource will be divided if cooperation takes place. We call this individual the "decision maker" and the other individual the "partner." Because we are testing the effects of partner choice, we allow the partner to decide whether or not to cooperate with the decision maker based on reputation. For simplicity, however, we do not model the formation of this reputation. We just assume that, for a reason we are not interested in here, the partner knows the offer that the decision maker would make if cooperation were to take place. If the decision maker's reputation is acceptable to the partner (see the next section for how individuals make decisions), the two individuals start to cooperate to produce the resource until the end of their interaction, which occurs at a constant split rate τ. The resource, which is equal to the sum of the productivities of the two individuals, is then shared according to the accepted offer. Conversely, if the decision maker's reputation is not good enough for the partner, the two individuals do not cooperate, and they return to the population without receiving any benefit. They can then meet new individuals at the same constant rate β.

Decision functions

Individuals need to be characterized by two decision functions to implement the social life described above: one function to produce offers, and one function to produce requests (the minimum offer that individuals are willing to accept when they play the role of partner). Both functions are genetically encoded and take as inputs the two individuals' productivity or effort. Here we detail how individuals react to individuals' productivity only; explanations regarding effort are the same.

Our goal is to simulate two cases: first a simple case in which only two productivities coexist in the population, and then a more general case where a continuum of productivities coexist. The level of complexity of the decision function required in the two cases differs, so we present them separately.

Two levels of productivity

With only two levels of productivity in the population, we assume that individuals are characterized by eight genetic variables: four p_{ij} and four q_{ij} variables, with i and $j \in \{HP, LP\}$ denoting an individual's productivity (HP = High-Productivity, LP = Low-Productivity). p_{ij} is the reputation (and offer) of a decision maker of productivity i in an interaction with a partner of productivity j. q_{ij} is the request, the minimum offer that a partner of productivity i is ready to accept in an interaction with a decision maker of productivity j. For example, a HP individual who is the decision maker in an interaction with a LP partner will have a reputation of p_{HPLP}. The LP partner will then compare the value of p_{HPLP} to his own q_{LPHP}, and if $p_{HPLP} \geq q_{LPHP}$, the interaction will be accepted and cooperation will start.

A continuum of productivities

Although introducing a continuum of productivities is necessary to get closer to biological reality, it constitutes a real challenge for modeling in that individuals need to be equipped with an infinity of traits to adapt their offers to the infinity of possible contributions by their partner (Gavrilets and Scheiner, 1993). To solve this problem, individuals now make decisions with two three-layer feedforward neural networks (one to produce offers, and another one to produce requests). Neural networks have been used previously to represent reaction norms in evolutionary models (Arak and Enquist, 1993; Enquist and Arak, 1994; Ezoe and Iwasa, 1997), but not, to our knowledge, in models of cooperation.

Both neural networks have the same structure: two input neurons, respectively sensing the individual's own productivity and that of the individual's partner; five hidden neurons; and a single output neuron which generates either the offer or the request, depending on the network. Each network has its own set of genetically transmitted synaptic weights (see Fig. 8A and SM3 section 1.3.1 (SM, 2015)), and evolution operates on these weights. Because evolution does not take place directly on offers or requests, individuals can evolve a reaction norm.

Finally, we also develop simulations in which individuals can differ through both their effort and productivity at the same time. In this situation, neural networks have four input neurons, representing an individual's own productivity and effort,

and his partner's productivity and effort. The functioning of the rest of the network remains the same as described above.

The cost of partner choice.

The cost of partner choice is implicit in the above model. It is a consequence of the time it takes to find a new partner after the rejection of an offer. Hence, the cost and benefit of being choosy are not controlled by explicit parameters, but by two parameters that characterize the "fluidity" of the social market: the "encounter rate" β, and the "split rate" τ. When $\frac{\beta}{\tau}$ is large, interactions last a long time (low split rate τ) but finding a novel partner is fast (high encounter rate β), and individuals thus should be picky about which offers they accept. On the contrary, when $\frac{\beta}{\tau}$ is low, interactions are brief but finding a novel partner takes time, and individuals should thus accept almost any offer.

Reproduction

We model a Wright-Fisher population with non-overlapping generations. Each generation lasts for a constant number of time steps, at which point all individuals reproduce according to the amount of resources they have accumulated throughout their life, and then die. We initialize all simulations with a population of stingy and undemanding individuals, offering nothing to their partners when they are decision-makers and accepting any reputation when playing the role of partners. We then observe how offers and requests evolve across generations.

4.3.2 Analytical model.

We develop an analytical model to study the simple case where individuals differ by their productivity (but not effort), and where only two productivities coexist in the population. The analytical model incorporates all of the features of the simulations presented above, but with one simplification: we assume that the total number of interactions accepted per unit of time is the same for each individual. With this assumption, rejecting an opportunity to cooperate does not compromise the chances of cooperating later, but on the contrary grants new opportunities. This situation is analogous to the condition where $\frac{\beta}{\tau}$ tends towards infinity in the simulations: social opportunities are plentiful at the scale of the length of interactions. The analysis of this model is presented in details in SM3 section 2 (SM, 2015).

4.4 Results

We first assume that all individuals invest the same effort into the production of the common good but can be of one of two productivities: high-productivity individuals are able to produce twice as much benefit as low-productivity individuals. At the evolutionary equilibrium, simulations show that low-productivity individuals give 66% of the total resource produced to their high-productivity partners when partner choice is not costly (Fig 6A, upper-right panel, circle markers). This offer is exactly proportional to the relative contribution of each partner, as high-productivity individuals produce 66% of the total shared resource. Similarly, when high-productivity individuals make offers to low-productivity individuals, they only propose 33%, an offer which low-productivity individuals accept, as it corresponds to their relative contribution (Fig 6A, lower-left panel, circle markers). Finally, both high-productivity and low-productivity individuals offer each other exactly 50% of the total resource when they meet as a pair, reflecting the fact that proportionality means equal division when contributions are equal (Fig 6A, upper-left and lower-right panels). This pattern of offers is confirmed by the analytical model (dashed lines in Fig 6A, and see SM3 section 2 (SM, 2015)).

Proportional divisions are also found when contributions differ not by the amount of benefits produced but by the time invested in cooperation (Fig 6B). Individuals who invest half as much time as their partner offer their partner 66% of the total resource at the evolutionary equilibrium. Conversely, individuals who invest twice as much time make offers of 33% to their partner, showing that the fitness-maximizing strategy in this situation is to make offers proportional to each partner's relative time investment.

Figure 6 also shows that partner choice is a key requirement for proportional offers to evolve. When we decrease the $\frac{\beta}{\tau}$ ratio, individuals spend more time looking for new partners and thus the cost of changing partners is increased. In this situation, offers remain very low over generations and never rise toward proportionality, regardless of differences in productivity (Fig 6A, triangle markers) or effort (Fig 6B, triangle markers). Figure 7 shows the distribution of offers made by low-productivity individuals to high-productivity individuals at the end of an 8,000-generation simulation, for different values of the $\frac{\beta}{\tau}$ ratio. Proportional offers of 66% can only evolve when $\frac{\beta}{\tau}$ is large, showing again that without partner choice, proportionality cannot evolve.

With a continuum of productivities in the population, offers and requests at the evolutionary equilibrium are proportional to contributions. Each individual who

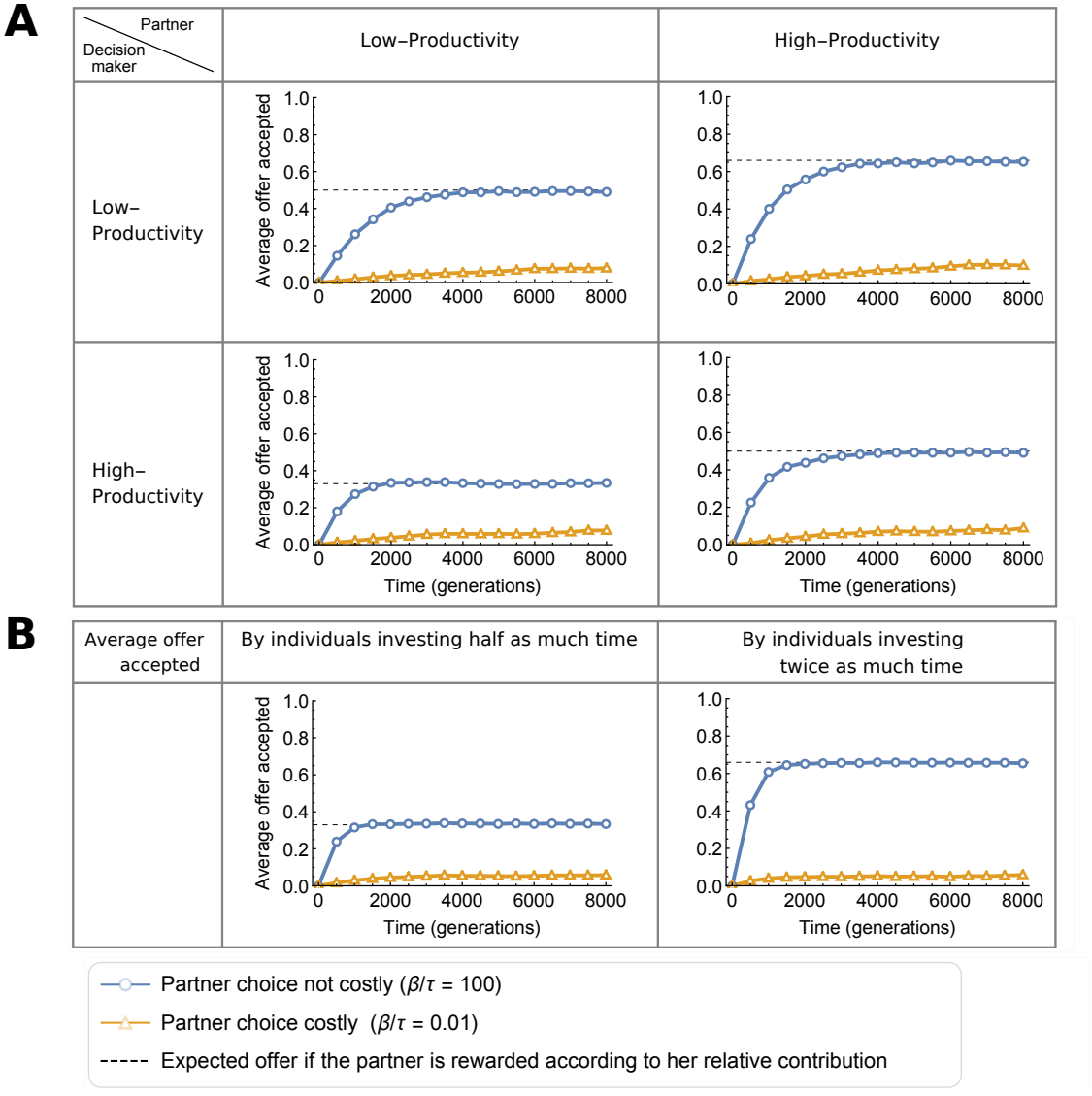

Figure 6: Evolution of the average offers accepted in cooperative interactions. **A:** Average offer accepted according to the productivity of the decision maker and the partner. High-productivity individuals produce twice as much resources as low-productivity individuals. When partner choice is not costly, offers evolve to match the partner's relative contribution. Dashed lines also represent the expected offer in the analytical model. **B:** Average offer accepted depending on whether partners invest twice as much or half as much time into cooperation. Individuals investing twice as much time receive twice the resources at equilibrium, and vice-versa.

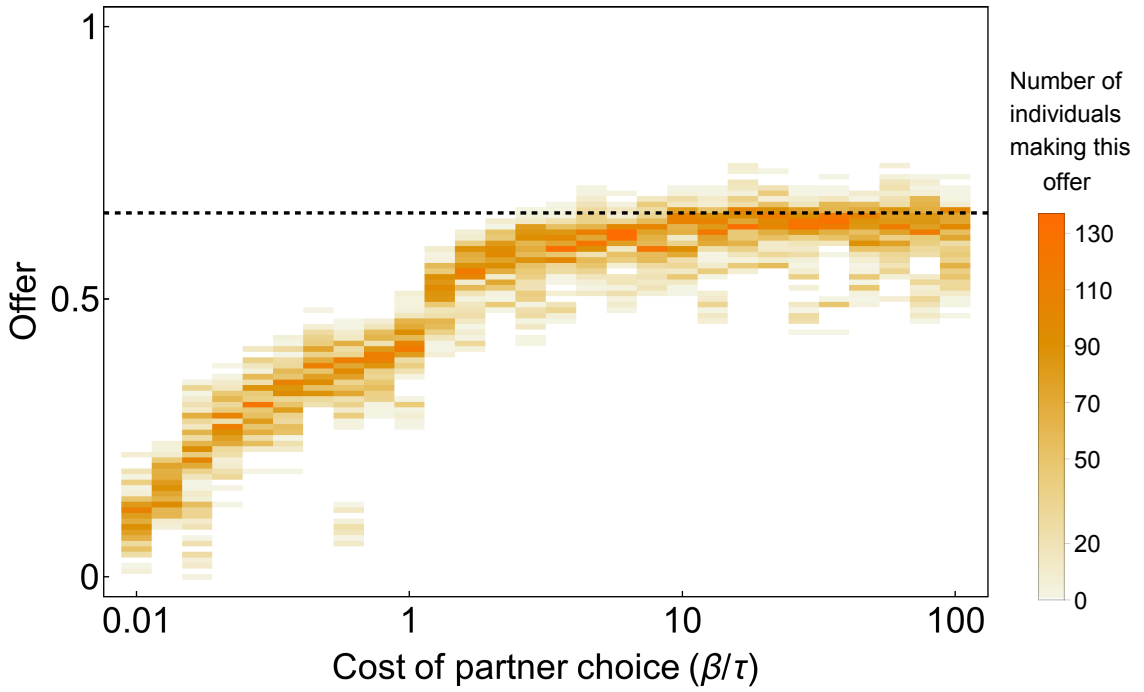

Figure 7: Distribution of offers made by low-productivity individuals to high-productivity individuals in the last generation of an 8,000-generation simulation, for different levels of partner choice cost (higher values of $\frac{\beta}{\tau}$ represent lower costs). High-productivity individuals' relative contribution compared to low-productivity individuals is 0.66, so the dashed line represents the expected equitable distribution. This distribution can only be reached when partner choice is not costly ($\frac{\beta}{\tau}$ is high).

enters an interaction is rewarded with an amount of benefits exactly equal to her productivity (Fig 8B). Plotting the average output of 15,000 neural networks after 8,000 generations of evolution shows that the networks evolved to produce offers and requests that are proportional to their bearer's relative contribution (Fig 8C and D, and see SM3 section 3.2 (SM, 2015)).

Finally, we performed a set of simulations to determine whether neural networks can take into account productivity and effort at the same time. The results are less clear-cut due to the increased number of input neurons, but they still show a trend toward greater offers to greater contributors (SM3 Fig 1 and SM3 section 3.2 (SM, 2015)), regardless of whether the differential contribution is due to differences in productivity or in time invested.

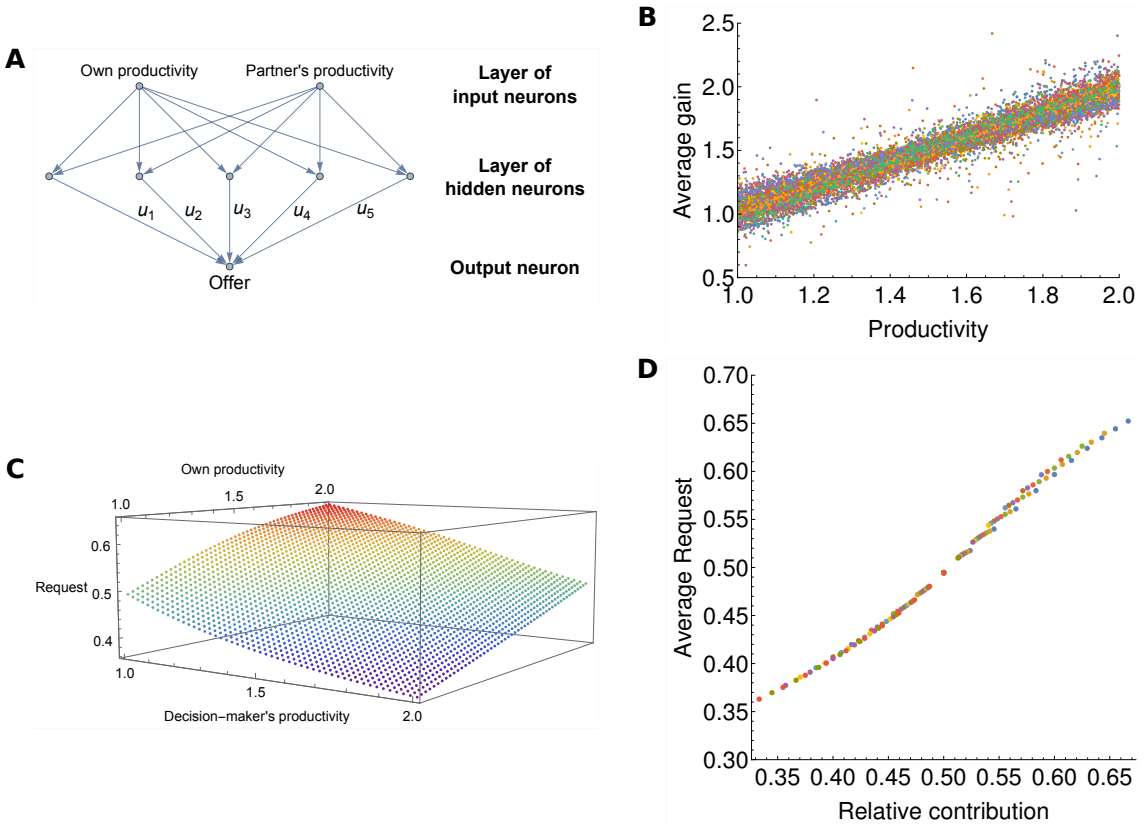

Figure 8: Evolution of equitable offers made by neural networks working on a continuum of productivities. **A:** Schematic representation of the neural networks that make offers. Networks take each partner's productivity as inputs and produce the offer as output. The u's represent synaptic weights on which evolution takes place. **B:** 15,000 individuals and their lifelong average gain plotted against their productivity. **C:** Average requests produced by the neural networks of 15,000 individuals after 8,000 generations, for different values of the input neurons. The more an individual produces and the less the decision maker produces, the larger the individual's request. **D:** Average requests produced by 15,000 neural networks plotted against the relative contribution of the bearer of the network.

4.5 Discussion

We modelled a population of individuals choosing each other for cooperation. When different contributions to cooperation are made, resource divisions proportional to contributions evolve. Refusing divisions that do not match one's own contribution and not proposing such divisions to others is the best strategy in such an environment. In other terms, a preference for equity maximizes fitness in an environment where individuals can choose their cooperative partners. This is true regardless of whether contribution is measured in terms of time invested or productivity.

It is important to note that our results cannot be summarized simply as "a preference for equity helps individuals to be chosen as a partner" or "a preference for equity helps avoid interactions with selfish partners." This is only half of the story. If the point were only to be chosen as a partner, the best strategy would be to be as generous as possible, an outcome which is sometimes observed in models inspired by competitive altruism theories (Roberts, 1998). The point here is rather to be chosen as a partner while at the same time avoiding exploitation by being over-generous. Our model clearly shows that the best strategy to solve this problem is to give proportionally to the other's contribution—not less, but also not more. Equity is the result of a trade-off between two evolutionary pressures which work in opposite directions.

This last point is better understood by looking at the precise mechanism through which proportionality evolves. The key factor determining divisions of resources at the evolutionary equilibrium are the outside options available to each individual. When assessing an offer, individuals are better off accepting if they could not get a "better deal" elsewhere. High-productivity individuals get more in our model because they have better outside options. Suppose that low-productivity individuals produce 1 unit of a resource whereas high-productivity individuals produce 2. High-productivity individuals thus have the option of producing 4 resources when they interact with other high-productivity individuals, leaving them with 2 resources on average (see exactly why in SM3 section 3.1 (SM, 2015)). Thus, if low-productivity individuals want to interact with them, they will have to give them exactly 2 resources (out of 3 produced), which will result in a proportional offer of 66%. But they should not give more, because they also have access to other interactions in which they could gain 1 unit on average. Each individual wants to make sure that they receive the same return on investment in each interaction that they enter into: each unit of time invested with a low-productivity individual should lead to as much benefit as each unit of time invested with a high-productivity individual.

Our model has several limitations, which need to be acknowledged. First, while

we suppose that individuals choose each other based on their reputation, we do not explicitly model the formation of this reputation. Individuals automatically know the reputation of others and this reputation is reliable. It could be interesting to relax this assumption, especially because reputation formation (through communication for instance) might be an important point that distinguishes humans from non-human primates. Second, the population we model does not match the hunter-gatherer population in the sense that it is not structured. This is important because a structure, such as camps or family units, could potentially affect opportunities to choose partners. Finally, it might be interesting to model the evolution of fairness in a wider range of cooperative interactions than we have considered here (outside distributive situations for instance). All of these assumptions should be relaxed in future studies.

Partner choice is not the only evolutionary mechanism postulated to lead to the evolution of fairness in the literature. Some authors have argued that fairness could be explained by empathy (Page and Nowak, 2002), spite (Huck and Oechssler, 1999; Barclay and Stoller, 2014; Forber and Smead, 2014), "noisy" processes such as drift or learning mistakes (Gale et al., 1995; Rand et al., 2013), the existence of a spatial population structure (Page et al., 2000; Killingback and Studer, 2001), or alternating offers (Rubinstein, 1982; Hoel, 1987). But as we explained in the introduction, all of these models equate fairness with equality, and it is thus unknown whether they can explain a more general case. Testing whether those models pass the "equity test" will be an excellent way to compare and decide between these models, a necessary undertaking that has been largely neglected.

Adopting a vocabulary of "offers" and "requests" suggests alternative interpretations of our model. It might be argued that human fairness is the result of bargaining at the proximal level, the result of rational cognitive processes. We argue instead that the "bargaining" already took place at the ultimate level by means of natural selection, and that the result of this bargaining is the existence of a genuine sense of fairness which "automatically" makes humans prefer equitable strategies. This hypothesis does not exclude the possibility that humans are also capable of consciously bargaining based on their outside options, but this behavior would not be the product of an evolved sense of fairness. This interpretation is well supported by empirical data showing that people do not justify their fairness judgments based on their outside options. For instance, in the famous ultimatum game, fairness judgments are clearly not based on outside options: low offers are rejected even when there is no available alternative (Güth et al., 1982; Camerer, 2003). Twelve months old children also react to inequity (Schmidt and Sommerville, 2011; Geraci and Surian, 2011), which can not be explained by conscious bargaining. But even

for adults, determining one's outside options in day-to-day life, across a vast range of situations that have never been encountered before, would seem to be a very complex and cognitively costly task (Chase et al., 1998; Todd and Gigerenzer, 2000). While our model bears a great resemblance to historical market models (Osborne and Rubinstein, 1990) and other models in economics in which fair outcomes have sometimes been observed (Rubinstein, 1982; Binmore, 2005), we emphasize that the markets we model are ultimate biological markets (Noë and Hammerstein, 1994; Noë et al., 2001). This is not just an empty terminological variation: locating markets at the ultimate level has important implications for our understanding of the psychological mechanisms underlying fairness. Among other things, it allows us to understand why fairness does not seem to be based on self-interest at the psychological level even if fairness evolved for self-interested reasons (Baumard et al., 2013; Trivers, 1971).

Another alternative interpretation of our model remains. One could agree that fairness judgments are based on simple automatic rules rather than complex conscious calculations, but argue that those rules could have evolved culturally rather than biologically. This is not an issue that can be settled theoretically, as the same models can always be interpreted as instances of biological or cultural evolution. To date, we definitely lack empirical data to answer this question with certainty, but the idea of a biologically evolved sense of fairness is not made absurd by the existing data. As early as the age of 12 months, children react to inequity (Schmidt and Sommerville, 2011; Geraci and Surian, 2011; Sloane et al., 2012), equity has been identified in many cultures around the word (Marshall et al., 1999; Gurven, 2004), and children reject conventional rules when they violate principles of fairness (Turiel, 2002). We do not take experiments on inequity aversion in non-human primates as evidence for a biologically evolved sense of fairness, as the negative reactions to inequity observed so far can still be interpreted in more parsimonious ways (see Bräuer and Hanus (2012) for a review and Amici et al. (2014) for methodological issues). Nonetheless, those experiments remind us that many researchers expect that prosocial behaviors traditionally associated with the existence of human institutions, religions, or cultural artefacts can also evolve biologically. In fact, Robert Trivers himself recognized that the most important implication of his seminal paper on the evolution of reciprocity (Trivers, 1971) was that "it laid the foundation for understanding how a sense of justice evolved" (Trivers, 2006).

The existence of intercultural and interindividual variations in fairness judgements (Henrich et al., 2005; Cappelen et al., 2010; Schäfer et al., 2015) is sometimes taken as evidence against their biological origin. This criticism is generally ill-founded, as evolutionary explanations have no particular difficulty accommodating

variation (Barkow et al., 1992). In the case of fairness, it is important to remember that what our model predicts is not the evolution of a fixed judgement but the evolution of an algorithm, an information-processing mechanism (Barkow et al., 1992). This is particularly evident in our extended simulations where the evolving unit is a neural network, precisely a special type of algorithm. This algorithm works on inputs (contributions) to produce outputs (divisions of resources), and here lies an important source of variability, because inputs can vary across cultures and individuals while the algorithm remains the same. For instance, measurements of contributions are affected by beliefs ("How long do I think it takes to harvest this quantity of food?"). If contribution was the only input in our model, in real-life more parameters can affect the algorithm's inputs, such as general knowledge ("Is this person not productive because she is sick?") or individual interpretations of the situation ("Are we engaged in a communal interaction? A joint venture? A market exchange?"). This last point could explain why even in carefully controlled environments, where there is little ambiguity about the source of inequalities, there is still heterogeneity in fair behaviors, with some people behaving as egalitarians, others as meritocrats, and others still as libertarians (Cappelen et al., 2007, 2010).

In fact, while interindividual and intercultural variations have crystallized the debate, intra-individual variation can also be observed even in Western countries. In some situations we behave as meritocrats, requiring pay for each additional hour of presence at work (Adams, 1963; Adams and Jacobsen, 1964), whereas the next day on a camping trip with strangers we behave more like egalitarians, without constant monitoring and bookkeeping of our contributions and those of others (Cohen, 2009). Neither our brain (the algorithm) nor our culture has changed in the meantime. What has changed is the way we interpret the situation (part of the input to the algorithm). This idea needs to be developed more formally, and we do not suggest that it is the only way to explain variation, but it may constitute a fruitful avenue of research.

Another interesting question is the prevalence of equity in traditional societies. We have mentioned anthropological records of distributions according to effort (Gurven, 2004; Kaplan and Gurven, 2005), but it is also well known that hunter-gatherers transfer meat in a way that not does not seem to respect equity. This type of interaction has been called "generalized reciprocity" by Sahlins (1972) and also seems to match Fiske (1992)'s notion of a "communal sharing" system. There are at least two mutually compatible ways to reconcile this observation with the predictions of our model. The first is to recognize that equity can be limited by other factors, for instance diminishing returns to consumption (Nettle et al., 2011). People could stop caring about equity when they become satiated or when they receive little additional

value from consuming one more unit of benefits. The second is to consider that even in generalized reciprocity good hunters are rewarded with more benefits, but those benefits are delayed. This hypothesis has received support recently from findings showing that generous hunters and hard workers are central in the social networks of small-scale societies (Lyle and Smith, 2014; Bird and Power, 2015). In this last perspective, our model should not be taken at face value as predicting the evolution of strict equity with immediate input/output matching, but more generally as input/output matching over a long time and across different cooperative activities ("generalized equity").

We conclude by noting that proportionality is important in distributive justice but is also a cornerstone of institutional justice, wherein offenders are punished in proportion to the severity of their crimes (Hoebel, 1954; Robinson and Kurzban, 2007). It is also central to the morality of many religions, in which rewards and punishments are made proportional to good and bad deeds by supernatural entities or forces (Baumard and Boyer, 2013). Although this is only speculation at present, our results may thus also explain why historically recent cultural domains such as penal justice and moral religions insist on the principle of proportionality: retributive punishment and supernatural justice may reflect our evolved desire for proportionality.

Chapter 5

Evolving fair robots

"I am SELFISH."

AN34NLDIU309

5.1 Objectives and summary

Simulations and models presented so far have a few problems: they constrain the behavioral repertoire of our agents, they do not assess the importance of mechanistic problems, and they limit the evolution of fairness to very special situations: distributive situations. The objective of this chapter is to start addressing these problems with a new methodology: evolutionary robotics. Evolutionary robotics frees us from making decisions about every aspect of our agents' behavioral repertoire, makes mechanistic problems stand out, and discovers evolutionary solutions that humans would not have thought of.

This is still a work in progress. I have no results to present so far, but I hope to be able to do so for my PhD defence. At the moment, I am still in the process of modifying virtual environments used in robotics so that they match the more natural environments I am interested in. I also need to modify how robots reproduce and interact together, as if evolutionary robotics is inspired by biological evolution, it does not aim at reproducing it perfectly.

5.2 Evolutionary robotics

Evolutionary robotics is the field concerned with the creation of robots that are both autonomous (do not need a human operator) and adaptive (can change their behavior according to changes in their environment. As this definition is confusing for a

biologist, I will use the word 'plastic' instead to refer to the property that roboticists call 'adaptive'). Building such robots is still a challenge for humans (Bongard, 2013). Traditional roboticists use machine learning techniques to manually optimize robots' behavior, but machine learning is not very good at producing plastic behaviors. The force best capable to produce entities that are both autonomous and plastic is natural selection. Hence, evolutionary roboticists harness the power of natural selection to evolve robots in virtual environments before selecting and manufacturing only the best robots. The long-term goal of evolutionary robotics is to "obtain an automatic process able to design, and even build, an optimal robot given only the specification of a task" (Doncieux et al., 2011).

For simple experiments like the one we are interested in, a robot is typically constituted of a wheeled platform supporting a different number of motors and sensors. The number and type of sensors is decided by the experimenter. Sensors can react to proximity of objects, light, color on the floor, sound, etc... The robot's controller is generally a neural network, taking as inputs the activations of each sensor and producing as output the velocity of each motor (see Fig 9 for an example).

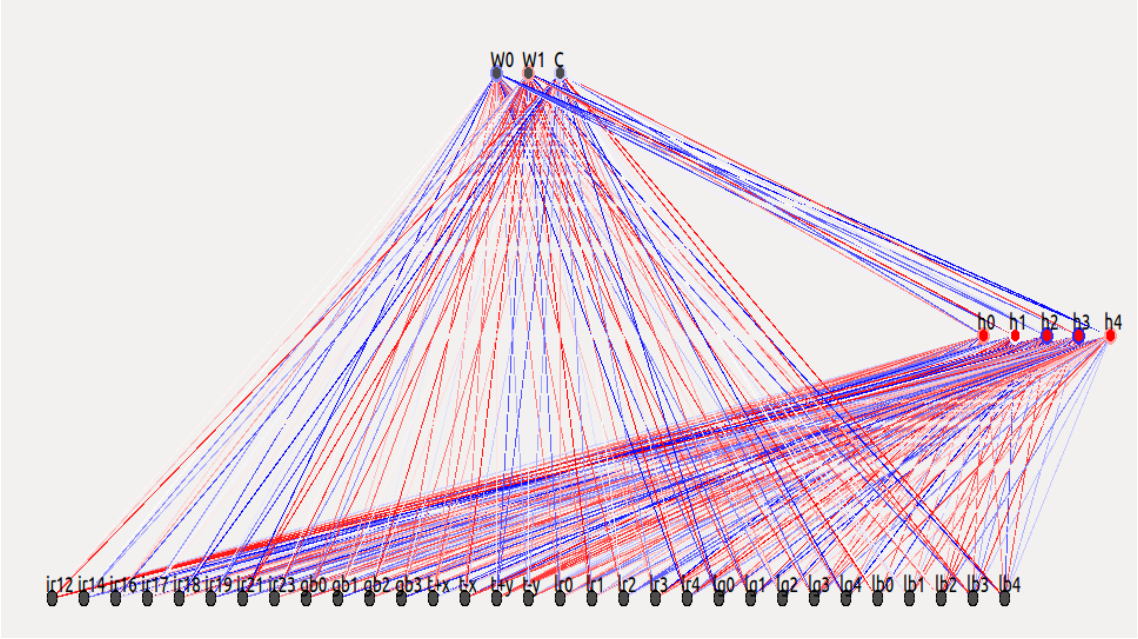

Figure 9: Neural network of the robots presented in Fig. 11. The bottom row represents the inputs, i.e. the activation states of all the sensors of the robot (in this case, proximity sensors, ground sensors, traction sensors, velocity sensors, and color sensors). The top row represents the outputs, the activation of the left (W0) and right (W1) motors, and the color of the robot (C). The middle row is a layer of hidden neurons allowing a more fine-tuned relationship between the input layer and the output layer. Although this network can seem pretty complicated, it can be optimized by natural selection without human intervention to produce the behaviors observed in Fig. 11. Image is a screenshot from the Farsa software (Massera et al., 2014).

As was the case in Chapter 4, evolution takes place on the synaptic weights of the neural network and the robot is rewarded for the behavior that the experimenter is interested in. For instance, imagine that the experimenter wants to evolve a robot that can detect a light source, head toward it and stay close to it. The experimenter generates a virtual environment in which a light source is present and parametrizes the simulator so that the robot is rewarded each time it is not further than x meter from the light source at the end of the simulation. A possible body plan for the robot is to be equipped with two motors and two light sensors (one on the right side, and one on the left side). Natural selection will then optimize the neural network so that the expected behavior is achieved consistently. For instance, natural selection could adjust synaptic weights so that when the left light sensor is activated, the right motor is more activated than the left motor, so that the robot turns left to head towards the light source (and vice-versa).

5.3 Advantages of evolutionary robotics

Evolutionary approaches have changed the way roboticists see the behavior optimization problem in many branches of robotics (Bongard, 2013). For us, evolutionary robotics has three important advantages:

- it frees us from making decisions about every detail of our agents' behavioral repertoire.
- it makes evolutionary mechanistic problems stand out
- it discovers solutions that a human investigator would not have thought of[1].

The simulations I have presented in Chapter 3 and 4 are highly constrained. Individuals always interact through an interaction that looks like an ultimatum game, they meet other individuals effortlessly, they know about their partners' reputation automatically, etc. We assume that individuals have a very limited behavioral repertoire, but that they excel at producing some of these behaviors. This is problematic because we might be overlooking problems - and in particular mechanistic problems - that could prevent the evolution of fairness in less constrained environments. In other words, we do not know how robust is the evolution of fairness without the behavioral repertoire coded in our simulations.

Analytical models are not better than simulations in this regard. Game-theoretical models have been criticized for making implicit but strong assumptions about the genetic structure underlying cooperative behaviors (André, 2014). For instance, jumping from a non-cooperative strategy to a cooperative strategy is often assumed to be easy (the matter of only one mutation). Many models are not interested in the mechanisms underlying the evolution of cooperation, and in particular genetic constraints. It is thus possible that many models are too optimistic at predicting the evolution of prosocial behaviors because they do not take those constraints into account.

Using evolving robots should help us to deal with those problems. With robots, we do not have to specify the exact design of the robot, and hence we do not have to specify the behaviors of the robot. It is in theory possible to allow the robot's body plan to evolve, but because we want to evolve a behavior that does not require

[1]For instance, in one of the first experiments in evolutionary robotics (Floreano and Mattiussi, 2008), a robot equipped with a camera had to learn to move toward certain shapes but not others. A human designer would probably try to solve this problem by taking whole pictures of the environment and analyze them to identify shapes, but surprisingly, the robot evolved to take into account only two small pixel patches out of the entire video stream.

a complicated body plan we could only allow the controller (the neural network) to evolve. We will thus have to determine what sensors our robot needs, but once this is done the behavioral repertoire will be potentially unlimited. For instance, in one of our experiments, we plan to replicate the evolution of fair divisions of resources (see below section 5.4). The difference this time is that robots will have to find by themselves the best way to divide the resource equally, rather than using an ultimatum game. Will the two robots necessarily consume the resource at the same time and at the same rate? Will one of the robots consume 50% of the resource at once before leaving access to the resource to her partner to consume the remaining 50%? Or will natural selection select for alternate consumption in small bouts in a way that will remind reciprocal grooming in impalas (Hart and Hart, 1992)?

Other mechanistic problems could become apparent. Maybe sharing in two equal halves is not the most difficult problem to solve after all, and the real evolutionary problem is to obtain a reliable reputation of other individuals. Observing the solutions favored by natural selection in unconstrained environments should teach us a lot about the practical problems related to the evolution of fairness.

5.4 Evolving fairness in distributive situations

The first thing I would like to use robots for is to replicate the simulations I have already done on the division of a resource. The goal would be to create a virtual environment strongly inspired by real environments, with robots who need to collaborate to find food patches (see Fig. 10). We will observe how robots divide those food patches at the evolutionary equilibrium when they can choose who they want to collaborate with. Because the food patches will be materialized and their size will decrease as consumption increases, it should be possible to visually assess how much each robot eats in a very natural way.

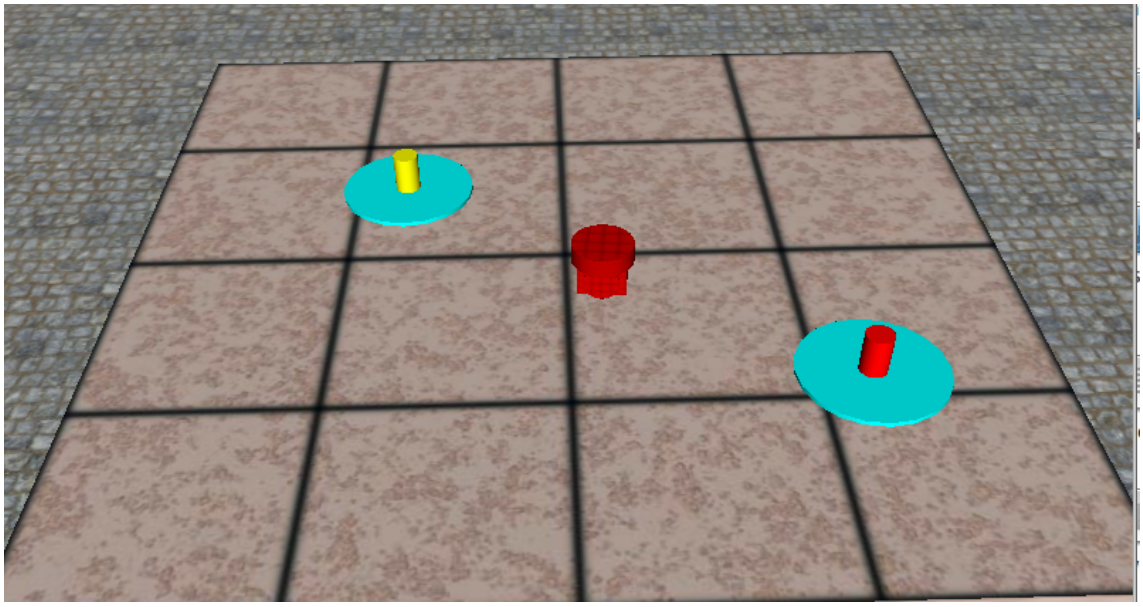

Figure 10: A robot (in the center in red) foraging for food patches (in blue). When a robot stands on a food patch, the robot's fecundity increases, and the patch diameter decreases. The goal is to build an environment in which many robots will need each other to find the food patches, and observe how they will share them at the evolutionary equilibrium. Image is a screenshot from the Farsa software (Massera et al., 2014).

5.5 Evolving fairness in situations of investment

As I stated in the introduction, fairness in the scientific literature is often linked to distributive situations. It is also the meaning I have adopted so far. But if the adaptive function of fairness is really to share the costs and benefits of cooperation impartially, there is no reason for fairness to be limited to the division of a resource. The amount of effort someone puts into cooperation is also a cost and should thus be taken into account. If fairness made people choose cooperative partners only based on how they divide resources, they would be exposed to cheaters who invest nothing into cooperation but agree to share equally.

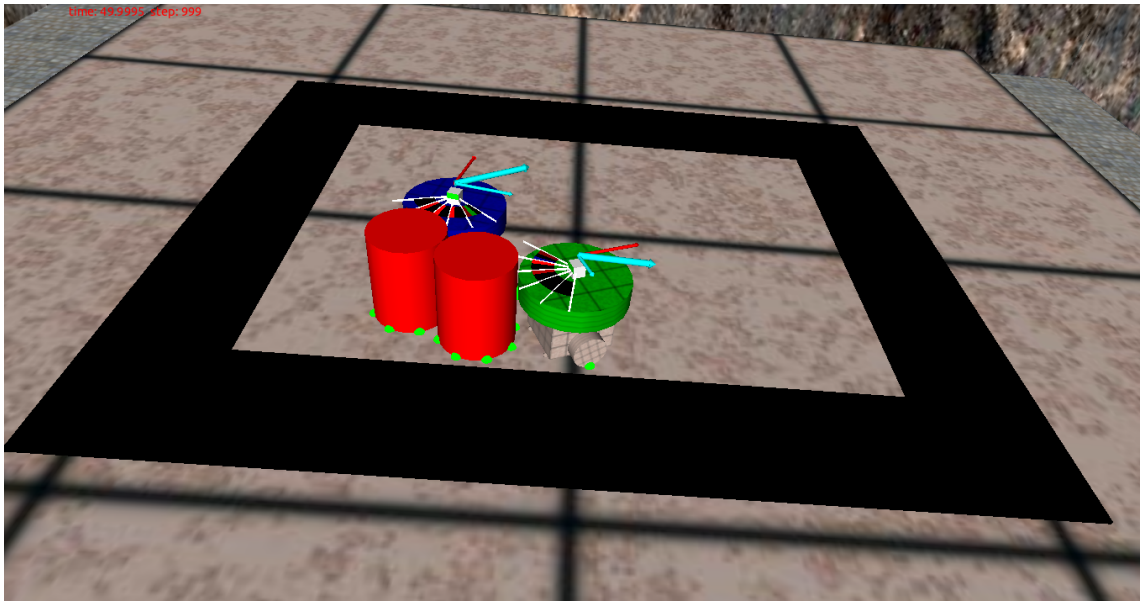

Figure 11: Two robots (in blue and green) cooperating to push a red object out of their arena. The robots' fecundity increases when the object gets out of the arena but the object is too heavy to move for a single robot - two robots need to collaborate to succeed. Again, the goal is to observe how fair will be the investment into pushing the object when robots can choose their cooperative partners. Image is a screenshot from the Farsa software (Massera et al., 2014).

Evolutionary robotics will thus be a perfect occasion to start addressing the evolution of fairness outside distributive situations (see Fig.11 for an example). There could be two interesting situations to distinguish. First, a situation in which there is a fixed amount of investment to be made to get the benefits of cooperation. If the benefits are shared equally, we should expect fairness to mean an equal investment from each partner. But there are also other situations in which the amount of investment is not predefined. For instance, when two people hunt together, they don't know how much time they will have to spend to find a prey. If one of the hunters decides to spend half of his time sleeping, the other hunter could find the situation unfair: he is paying opportunity costs because he could have chosen another partner to cooperate with and obtained more benefits this way. In this situation, we could expect that the fair investment would be to make as much effort as possible (maybe in the limits of one's capacities or one's partner's investment).

Part IV

Testing the partner choice theory empirically

Chapter 6

Partner choice creates fairness when outside options are equal (Paper 4)

"Because I wanted to maximize my profits after all isn't that what everyone does anyways they try to maximize their own profit so I was in a position so I took it."

ALD82N12NCL

6.1 Objectives and summary

The effect of partner choice on human behavior has already been investigated experimentally in many studies (Barclay, 2004, 2006; Barclay and Willer, 2007; Chiang, 2010; Sylwester and Roberts, 2013). All those studies support the theory of competitive altruism (Roberts, 1998), which predicts that when people are in competition to be chosen as social partners, they will increase their level of helping/generosity/cooperativeness. An interesting question that has received less attention is up to what point should people compete to appear generous (but see Barclay (2011, 2013)). Being fair is different from being as generous as possible, so we thought it would be interesting to experimentally show how partner choice can lead to the evolution of fairness, as opposed to the evolution of over-generosity.

As explained in previous chapters, the key is to manipulate outside options on which partner choice operates. It was not that easy to imagine how to create an experiment that would allow us to manipulate outside options in an ecologically meaningful way. We decided to create small groups of four subjects in which some subjects, the proposers, had to make offers to other subjects, the receivers, as to how a sum of money should be divided. This bargaining process was repeated over many

rounds. To evolve fairness, the trick was to enable receivers to become proposers in the next round if they were not satisfied of the offers they received, and, conversely, to allow proposers to become receivers if they were not chosen often enough. Because all subjects could play any role there was to play, they effectively had the same best outside option, and fair offers evolved. Allowing subjects to change roles was for us an (imperfect) way to reconstruct in the lab the type of environment with equal social opportunities for everyone that small-scale societies can present. Conversely, fixing the roles of proposers and receivers is probably adequate to represent situations in economics where the roles of sellers and buyers never change, but it did not seem realistic for us to postulate that a hunter-gatherer will always be stuck in the same social role in all his lifelong interactions.

The rest of this chapter comes from an article published in:

Debove, S., Andre, J. & Baumard, N. Partner choice creates fairness in humans. *Proc. R. Soc.B* 282, (2015).

Jean-Baptiste André built the analytical model presented in this article.

Partner choice creates fairness in humans

Abstract: Many studies demonstrate that partner choice has played an important role in the evolution of human cooperation, but little work has tested its impact on the evolution of human fairness. In experiments involving divisions of money, people become either over-generous or over-selfish when they are in competition to be chosen as cooperative partners. Hence, it is difficult to see how partner choice could result in the evolution of fair, equal divisions. Here, we show that this puzzle can be solved if we consider the outside options on which partner choice operates. We conduct a behavioral experiment, run agent-based simulations, and analyze a game-theoretic model to understand how outside options affect partner choice and fairness. All support the conclusion that partner choice leads to fairness only when individuals have equal outside options. We discuss how this condition has been met in our evolutionary history, and the implications of these findings for our understanding of other aspects of fairness less specific than preferences for equal divisions of resources.

6.2 Introduction

Partner choice is a major force that has driven the evolution of cooperation in humans (Barclay, 2013). Experimental studies show that in situations where people choose others as cooperative partners, individuals try to outbid competitors by increasing their investment in cooperation (Barclay, 2004; Barclay and Willer, 2007; Sylwester and Roberts, 2013). Investing more in cooperation is costly but also leads to a good reputation: if partner choice is possible, the benefits of being a good cooperator can outweigh its costs (Barclay, 2006; Sylwester and Roberts, 2010). Theoretical models point in the same direction: incorporating partner choice in models of cooperation selects for cooperative behaviors (Sherratt and Roberts, 1998; Aktipis, 2004; McNamara et al., 2008). All of these studies support the theory of "competitive altruism" (Roberts, 1998; Barclay, 2004): when people monitor and choose others on the basis of their cooperative behaviors, costly cooperative behaviors can pay off.

Although the importance of partner choice for the evolution of human cooperation is clear, very little is known about its importance for the evolution of fairness. Most studies on partner choice are primarily concerned with how much people invest in cooperation and not how people *divide* the common goods produced through cooperation. The most famous experimental evidence of fairness in the division of a good comes from the ultimatum game (Güth et al., 1982; Güth and Kocher, 2013). In this two-player laboratory experiment, one of the players (the "proposer") makes an offer to the other (the "responder") on how to divide a sum of money. If the offer is accepted then both players receive the money, otherwise none of the players receives any money. Traditional game theory, which assumes players to be super-rational, predicts that responders will accept any offer, however small, because getting something is always better than getting nothing. Anticipating this, proposers should only offer the smallest possible amount. But experimental tests have not confirmed these theoretical predictions: proposers' modal offer usually falls between 40 and 50%, and responders are prepared to reject very low offers just for the sake of "fairness" (Hagel and Roth, 1995; Camerer, 2003).

To our knowledge, the only evolutionarily-minded paper studying the impact of partner choice on the fairness of money divisions is a study by Chiang (2010). Using a repeated ultimatum game, Chiang (2010) shows that partner choice increases offers from 42.20% to 46.28%, getting closer to the "fair" expected offer of 50% after 15 repetitions. Chiang (2010) concludes that his findings are "consistent with the predictions of competitive altruism theory" which is interesting because the predictions of competitive altruism theory in this case have not always been thoroughly

discussed. Indeed, an interesting evolutionary question is to know up to what point people should attempt to appear generous when partner choice is possible. Some authors have argued that people will increase their generosity until the marginal costs of doing so exceed the marginal benefits, but what costs and benefits should be taken into account remains unclear.

Outside the evolutionary field, the consequences of partner choice for the evolution of fairness are studied in behavioral economics. In a seminal paper by Roth and Prasnikar (1991), nine proposers are in competition to make offers to a single responder. The responder then chooses an offer - and thus a single proposer. In this experimental setup, offers rose very rapidly to 99.5%. The same pattern of highly generous offers was replicated in Fischbacher et al. (2009), and a similar "runaway" effect of partner competition has been found in laboratory market experiments (Cason and Williams, 1990).

Interestingly, a few studies have showed that partner choice can also lead to the opposite pattern of offers - extremely selfish offers (Güth et al., 1998; Grosskopf, 2003; Fischbacher et al., 2009). In Fischbacher et al. (2009) for instance, two responders were in competition to access the offers made by a single proposer. After 20 repetitions of the game, the average offer decreased to 18.8%. The effect was even more dramatic when five responders were in competition to access the offer of a single proposer: proposers became increasingly selfish and offered an average of 13.8% in the last repetition.

In summary, a cross-disciplinary review reveals that partner choice leads to very unbalanced divisions of benefits in two opposite directions: the proposer either makes highly generous offers or highly selfish ones. In this paper, we aim to understand the origin of these opposite findings. We hypothesize that it is not partner choice in itself which is responsible for such unbalanced divisions, but rather unequal 'outside options.' Outside options are the individual's expected payoff in the same timespan if she had refused the current interaction. It is perfectly possible to be able to choose partners but only have bad options to choose from: in this case, it will be difficult to know whether unbalanced divisions are the result of the mere possibility to choose partners or of the existence of those bad outside options. We predict that when partner choice is possible, players should be "rewarded" according to their outside options: if proposers have better outside options than responders, runaway selfishness should be the result. If responders have better outside options than proposers, runaway generosity should be the result. Finally, and more importantly, we hypothesize that when proposers and responders have the same outside options, partner choice leads to a fair, 50/50 division.

We tested this hypothesis empirically and theoretically. In the behavioral experiment, groups of four participants played a modified version of the dictator game that allows for partner choice. We contrasted a condition in which proposers had better outside options than responders to a condition in which responders had better outside options than proposers. We predicted that offers would be over-selfish in the first case and over-generous in the second. In a third condition, we equalized the outside options of proposers and responders and predicted that fair offers would evolve. In the agent-based simulations and the game-theoretic model, we considered larger populations of agents and introduced a continuum of partner choice to demonstrate the robustness of the evolution of fairness when outside options are equal.

6.3 Behavioral experiment

Methods

The experiments were conducted in March and May 2014 at the Nuffield Centre for Experimental Social Sciences (CESS). All conditions were approved by the CESS Ethics Review Committee. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007).

Participants

A total of 120 participants were recruited from the University of Oxford using a web-based recruitment system. Participants were told that they would earn £4 for showing up and would earn additional money during the course of the experiment. The average earnings per subject were calculated to be at least £10 per hour.

General Procedure

Participants were seated at computer terminals separated by partitions so that they could not see one another. We also ensured anonymity: the subjects did not have access to identifying information about the other players at any point during (or after) the experiment. Once seated, participants read instructions that explained the procedure. The instructions were then read aloud by the experimenter while participants read along. Participants then had time to ask questions. The participants first performed a practice round, followed by 30 experimental rounds. After the experiment, subjects answered a questionnaire about their behavior and thought process in the experiment.

The experiment included three conditions: competitive altruism (CA), runaway selfishness (RS) and equal options (EO). Each participant played in only one condition. Forty subjects took part in the CA condition, 44 in the RS condition, and 36 in the EO condition (differences are due to unequal show up rates between conditions). Conditions CA and RS present asymmetries of outside options between proposers and responders and should allow us to replicate the results of previous studies. They also serve as a baseline against which to compare results from the EO condition, in which outside options are equalized.

We start by describing the procedure common to all conditions before detailing procedures specific to each condition. In all conditions, subjects played 30 rounds of a four-player game. Groups of four were stable across rounds. At the beginning of each condition, subjects were randomly assigned to one of two roles ("proposer" or "responder"), and were informed of their role. In each round, proposers and responders could form partnerships to split a pool of money: proposers made offers, and responders could accept them. Subjects were informed that they would gain their average payoff across all rounds. Subjects did not know how many people were in their group, nor the number of proposers and responders in each round. The only information they had was whether or not one of their offers was accepted (proposers) or what offers remained available to them in the current round (responders). This enhanced the probability that the money divisions we observed in our experiment would result mechanically from each individual's outside options, and not from strategic thinking.

Conditions

Competitive Altruism (CA)

In the CA condition, there were three proposers and only one responder in each group of four subjects: proposers were thus in competition to be chosen by responders. Proposers and responders kept the same role for the entire 30 rounds. Each round proceeded in the following steps:

- Each proposer in a group decides what division of £10 with a responder from the group to propose.
- The one responder in the group chooses among the proposers' offers (with the obligation to choose one offer - she cannot refuse them all).

- Participants are informed of their own earnings for the round. The responder and the selected proposer receive the portions of the £10 corresponding to the chosen offer. The two proposers who were not chosen by a responder earn £0.

Because in this condition proposers are in competition to be chosen by responders, their outside options are worse than those of responders. We thus predicted that this asymmetry would lead to biased money divisions in favour of responders.

Runaway Selfishness (RS)

The RS condition is similar to the CA condition, except that the number of proposers and responders in each group was reversed: three responders were in competition to access the offers made by a single proposer. Each round proceeded in the following order:

- The only proposer in the group makes an offer
- One responder in each group is randomly selected to accept this offer (as in a dictator game, the responder cannot refuse it)
- Earnings are reported to each participant. The two responders who were not selected for the offer in this round earn £0 in this round.

In this condition, responders had worse outside options than proposers, because they were in competition to gain access to proposers' offers. We thus predicted that partner competition would lead to biased money divisions in favour of proposers and ever-decreasing offers - runaway selfishness.

Note that subjects who participated in the CA and RS conditions received the exact same sheet of instructions. The only difference between the two conditions was the number of proposers and responders in each group, a parameter that was not communicated to subjects. Hence, any difference in behavior observed between these two conditions can only be attributed to the change in this parameter, and the resulting difference in the asymmetry of outside options between the conditions. Note also that if subjects knew their number of rivals, we would make the same predictions. We decided to give subjects as little information as possible to increase the probability that the effects we could observe would be the result of a "mechanical" effect of outside options, and not the result of strategic thinking (although we can not entirely rule out this possibility).

Equal Options (EO)

In the EO condition, all subjects had the same outside options. Although the condition began with two randomly selected proposers and responders in each group, subjects could decide to switch roles at the end of each round after having been informed of their payoff. Hence, proposers and responders who were not satisfied with their payoff could decide to play the opposite role in the next round. As before, subjects were not informed of the current number of proposers and responders in their group, nor were they informed of how many people were willing to change their role in the current round. In case all four subjects decided to play the same role, no partnership was concluded in the next round and all subjects received a null payoff.

We predicted that because outside options were equal in this condition, partner choice would lead to a stable "fair" equilibrium and the evolution of equal divisions. Note that having the same number of proposers and responders in each group but with *fixed* roles would not be enough to make this condition different from the RS condition: with an equal number of proposers and responders, proposers see their offers being accepted in each round, and should decrease their offers as in the RS condition.

6.3.1 Results

Figure 12 plots the evolution of the average offer accepted in each condition. We include the first round in this graph for informative purposes, but this round was a practice round and was not included in statistical analyses. Figure 12 confirms our predictions: each condition influenced the offers in the expected direction. In all rounds except the first (practice) round, mean accepted offers $\bar{o}$ followed the inequality $\bar{o}_{CA} > \bar{o}_{EO} > \bar{o}_{RS}$. Figure 12 also suggests an increasing trend over time in the CA condition and a decreasing trend over time in the RS condition.

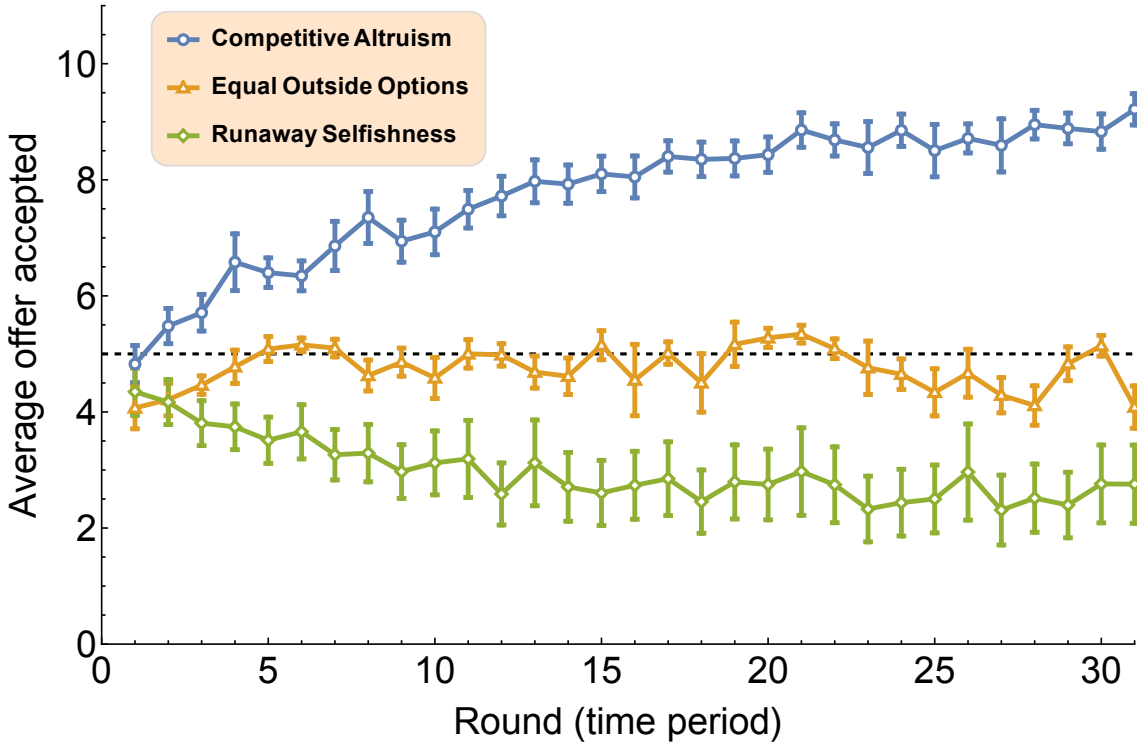

Figure 12: Evolution of the average offer accepted by responders in each of the three conditions. In the Competitive Altruism condition, responders have better outside options than proposers. In the Runaway Selfishness condition, proposers have better outside options than responders. In the Equal Outside Options condition, proposers and responders can choose partners and have the same outside options. Round 1 is a practise round. Error bars represent standard errors.

We tested the significance of the differences between conditions using Mann-Whitney tests. We analyze each group as an N of 1 as a way of dealing with non-interdependence of decisions within a group. Significant differences were found between all conditions in pairwise comparisons: CA and EO (n1 = 9, n2 = 10, U = 90, $p < 0.001$), EO and RS (n1 = 9, n2 = 11, U = 17, $p = 0.012$), and RS and CA (n1 = 10, n2 = 11, U = 110, $p < 0.001$). An nptrend test (Cuzick, 1985) rejected the null hypothesis that there was no trend across conditions ($z = 4.65$, $p < 0.001$). Using data from the last ten rounds only, the differences between pairs of conditions CA and EO, EO and RS, RS and CA remained significant ($p < 0.001$ and U = 90, p = 0.028 and U = 21, $p < 0.001$ and U = 110 respectively), and the nptrend test was still significant ($z = 4.55$, $p < 0.001$). Differences were still significant at least at the 5 percent level when the last eight or twelve rounds were analyzed instead of the last ten.

	All rounds	Last 10 rounds
Constant	4.669**(0.171)	4.375**(0.299)
RS	−2.366**(0.228)	−1.741**(0.398)
CA	4.779**(0.233)	4.622**(0.407)
time	−0.006(0.0102867)	−0.046(0.055)
time x RS	−0.036*(0.013)	0.060(0.074)
time x CA	0.115**(0.013)	0.095(0.075)
N	878	295
R^2	0.71	0.76
F	429.054	188.731
$Prob > F$	0.000	0.000

Table 3: Pooled regression predicting the average accepted offer. Reported numbers are ordinary least squares coefficients. Numbers between parentheses are standard errors. The left column gives a regression using data from all rounds. In the right column, only data from the last 10 rounds were used.

RS = Runaway Selfishness.
CA = Competitive Altruism
* = Significance at the 0.01 level
** = Significance at the 0.001 level

Table 3 shows the results of a regression analysis of the average offer accepted in the round, pooling the three conditions CA, RS and EO, and setting EO as the omitted category. In Column 1, all rounds are considered and numbered from -29 to 0 so that the reported coefficients in the table represent effects in the last round of the experiment. In Column 2, only data from the last 10 rounds is used, and rounds are numbered from -9 to 0. The data were checked for linearity, normality, homoscedasticity, and autocorrelation.

The estimated accepted offer in the last round of the EO condition was 4.67 (column 1, line 1), very close to the 50 % fair division. The negative coefficient in the RS condition and the positive coefficient in the CA condition, both significant,

show that the predicted offers in these conditions differed in the expected direction. Offers in the CA condition are expected to be 4.8 units higher than in the EO condition, and offers in the RS condition are expected to be 2.4 units lower than in the EO condition. A comparison of column 1 with column 2 shows that there is no substantial difference of average offers accepted in the last 10 rounds compared to all rounds, controlling for time trends.

Time did not have a significant effect on the offers accepted in the EO condition (column 1 and column 2): offers remained stable across all rounds in this condition. On the contrary, significant interactions were found between time and RS and time and CA. The effect of time was especially large in the CA condition: offers increased by 0.12 units at each round. However, these interactions were no longer significant in the last 10 rounds (column 2), suggesting that offers ended up reaching a stable level in all conditions, as is already suggested by Figure 12.

6.4 Theoretical model

Methods

We model a population of agents who have the same outside options and play ultimatum games repeatedly throughout their lifespan. Individuals meet each other in pairs at a constant rate β. When they meet, one individual is randomly selected to play the role of proposer, while the other plays the role of responder. The proposer makes a genetically encoded offer to the partner. If the offer is accepted, the two partners enter a cooperative interaction which is assumed to take time. During this cooperative interaction, they divide a resource of size 1 according to the accepted offer, until the end of the interaction which occurs at a constant rate τ. If the proposer's offer is rejected, the two partners are separated without interacting and return to the population to find an unpaired partner.

At the end of their life, all individuals reproduce according to the amount of resource they have accumulated. Individuals pass on their offers and requests (the minimum offer they are ready to accept when they play the role of responder) to their offspring, with a small probability of mutation on these traits. The model is fully explained in SM4 section 1 and SM4 section 2 (SM, 2015).

When the encounter rate β is high, it is easy to find a new partner in the population. When the split rate τ is low, interactions last a long time. Hence, when the $\frac{\beta}{\tau}$ ratio is high, partner choice is not costly, as rejecting an unfair offer does not mean that time will be wasted looking for a new partner. Moreover, because the

roles of proposer and responder are assigned randomly in each new encounter, all individuals have the same outside options. In this environment where all individuals have the same outside options and can choose their cooperative partners, we observe what offers are made at the evolutionary equilibrium, which represent the fitness-maximizing offers. We also produce a resident-mutant analysis of the model, allowing us to pinpoint the offers that cannot be invaded by mutants once they have spread in the population. This analysis is detailed in SM4 section 2 (SM, 2015).

Results

Our simulations show that the average offer accepted in the population tends toward 50% at the evolutionary equilibrium when partner choice is not costly (Fig. 13, plain lines). The resident-mutant analysis shows that a resident population cannot be invaded by mutants as long as the offer p characterizing the population lies in the interval:

$$p \in \left[\frac{\beta/2}{\beta+\tau}, 1-\frac{\beta/2}{\beta+\tau}\right] \tag{6.1}$$

Hence, when partner choice is not costly ($\beta >> \tau$), the range of evolutionary stable offers is restricted to $p \in \left[\frac{1}{2}, \frac{1}{2}\right]$. Analytical results are thus in perfect agreement with simulation results and confirm that partner choice in a context of equal outside options leads to the evolution of fairness.

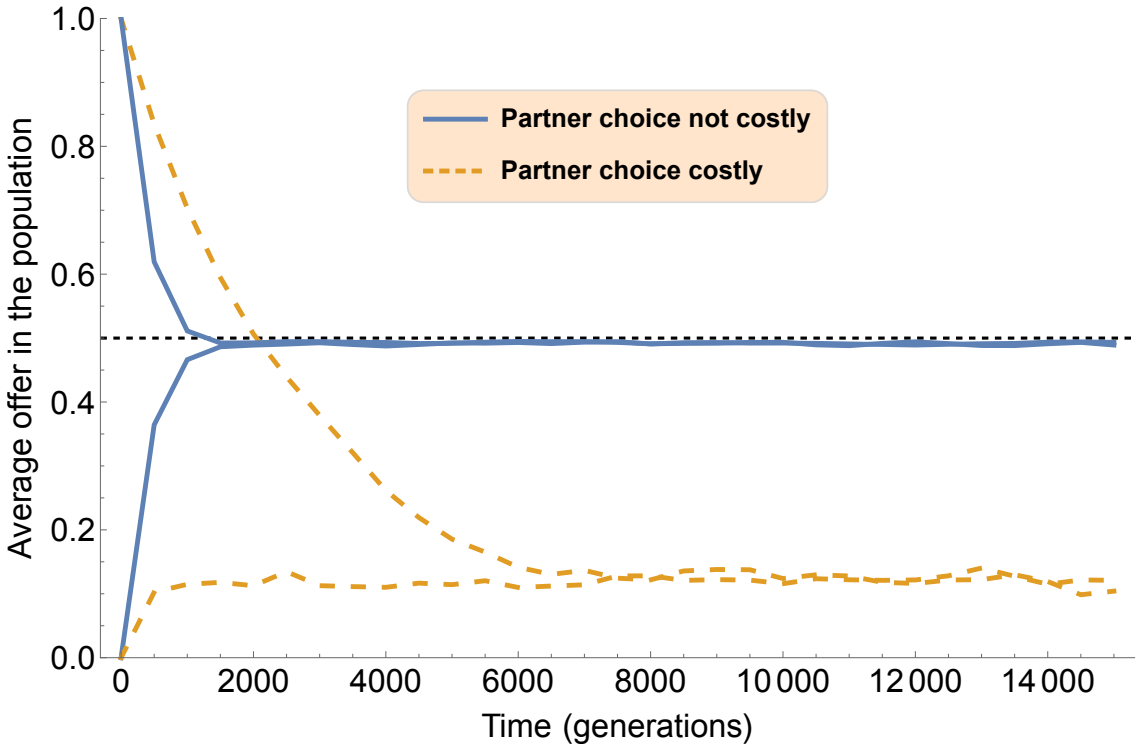

Figure 13: Evolution of the average offer in the ultimatum game when individuals have the same outside options and for two different costs of partner choice (simulation results). Two starting points (0 and 1) are used for each cost of partner choice. Each curve represents an average over 20 simulation runs. Parameter values used for these simulations can be found in SM4 section 2.2 (SM, 2015).

On the other hand, simulations show that when partner choice is costly ($\beta << \tau$), the average offer accepted at the evolutionary equilibrium is very low (Fig. 13, dashed lines): proposers can afford to be selfish. This result holds whether we consider an initial population of "over-selfish" individuals offering 0% or an initial population of "over-generous" individuals offering 100% of the resource to their partner, showing that our results are not limited by the initial conditions of our model (Fig. 13, dashed lines).

6.5 Discussion

Our study shows that partner choice creates fairness, but only in a context of equal outside options. Partner choice is the mechanism that allows individuals to receive offers corresponding to their outside options; whether or not the offers will be fair depends ultimately on the equality of those outside options. This emphasis on outside options also helps explain why previous studies reported opposite effects of

partner choice. Although the subjects in our CA and RS conditions received the exact same instructions, the inequality of outside options between the two conditions led to the evolution of offers in two opposite directions. Specifically, an asymmetry of outside options in favor of responders leads to runaway generosity, whereas an asymmetry in favor of proposers leads to runaway selfishness.

Although the importance of outside options may have been overlooked in evolutionary studies, it has already been investigated in behavioral economics (Cason and Williams, 1990; Knez and Camerer, 1995). A parallel could even be drawn between our results and the classical idea that an excess of supply or demand affects the price at which a commodity is exchanged (for a discussion of this parallel, see André and Baumard 2011b). Nonetheless, in behavioral economics, most studies fix outside options a priori and observe, once they are fixed, how they affect prices or bargaining outcomes. Here, on the contrary, we provide a condition in which equal outside options emerge endogenously from a partner choice-based environment. Hence, our work contains two main contributions to the literature. For scholars with a biological background, we draw attention to the prime importance of outside options when studying human partner choice and the evolution of fairness. And for scholars with an economics background, we show how the well-known effects of outside options are not limited to economic markets, but also have an impact over longer, evolutionary timescales, in "biological markets" (Noë et al., 1991; Noë and Hammerstein, 1995; Noë et al., 2001). In a nutshell, what we suggest is that the human sense of fairness is the result of natural selection optimizing human behavior in a market environment (without neglecting potential cultural or contextual effects, see Baumard et al. 2013).

Our work represents a number of methodological advances on previous related work. First, it uses a modified version of a dictator game, rather than an ultimatum game, to measure the fairness of money divisions: when only one offer is left, responders have no choice but to accept it. As divisions in the dictator game are known to be more asymmetric than those in the ultimatum game (Hagel and Roth, 1995; Camerer, 2003), the dictator game offers a more conservative way to observe the evolution of fairness. Second, we modified the dictator game so that it can be played not only between two players but in groups of four players, to introduce a first level of partner choice. A second level of partner choice is implemented by allowing subjects not only to choose their partner but also to change role between rounds. Finally, we observed behaviors on a longer timescale and with more independent observations than in previous studies.

The mechanism leading to fairness in our EO condition is easy to understand.

When there are more proposers than responders in a group, offers start to increase following the predictions of competitive altruism. But as offers rise, proposers start to receive decreasing payoffs, which leads some of them to decide to play responder in the next round. This incentive to switch roles in turn leads to an excess of responders over proposers. At some point, the asymmetry of outside options is reversed, and responders want to change role and become proposers. These two forces working in opposite directions lead to the evolution of fair, balanced divisions which oscillate around 50%. The mechanism at play is similar in our theoretical study: proposers cannot make offers lower than 50%, as responders would reject them and prefer to play proposer. Conversely, proposers have no incentive to make offers higher than 50%, as they would be better off playing responder themselves to benefit from those generous offers.

Although the mechanism in our study is clear, it is interesting to ask what its biological equivalent in the real world might be. The roles of proposer and responder are a convenient way to model asymmetries of bargaining power in the lab: the proposer is in a strategically advantageous position because the responder has no choice but to accept her offer. Allowing subjects to change roles means removing this asymmetry from the game. Although it is hard to imagine a strict equivalent of the roles of proposer and responder in nature, asymmetries of bargaining power are plentiful. For example, a physically stronger individual could benefit from a local competitive advantage at the moment of sharing the benefits of cooperation. Weaker individuals cannot "choose" to become stronger in this situation, so what could be the ecological equivalent of being able to change role from proposer to responder and vice versa? We suggest it is a way to implement the variety of roles humans play across all their lifelong cooperative interactions, including interactions in which they are not the weakest anymore. This assumption is well justified by the empirical literature on human cooperation: humans cooperate frequently and in diverse contexts, both with kins and non kins (Hill, 2002; Hooper et al., 2014). In a review of the human social organisation, Kaplan et al. (2009) insist on the "high-quality, difficult to acquire resources" hunter-gatherers consume, which require "high levels of knowledge, skill, coordination". Because knowledge, skill, or coordination are not necessarily correlated with physical strength, weak individual can be good cooperators and have access to good outside options even if they are locally in a poor bargaining position. In a sense, we think it is interesting to reverse the question: what could be the ecological equivalent of playing a repeated dictator game when the roles of proposers and responders are fixed? Whereas it can probably adequately represent some situations in economics where the roles of sellers and buyers never change, it does not seem realistic for a hunter-gatherer to always be stuck in the same

social role in all his lifelong interactions. Hence, without saying our paradigm is a perfect representation of humans' social life, we think it captures some interesting aspects of it, and is thus worth exploring.

Our study has a few important limitations. First, the small number of subjects in our groups means that the offers in each round may have been sensitive to noise. It would also be interesting to introduce asymmetries of outside options in a more natural way than by artificially fixing the number of proposers and responders in each group. In our theoretical study, we do not consider variations between individuals in outside options in the form of strong and weak individuals, for example. We also do not model the formation of reputation, as we suppose individuals have perfect information on the past behavior of other individuals. Examining if and how less-than-perfect information could prevent the evolution of partner-choice based fairness would thus be another way to extend our results.

Nonetheless, our study has interesting implications for our understanding of the evolution of human fairness as a whole. Whereas almost all theoretical studies of the evolution of human fairness have studied the evolution of equal divisions in the ultimatum game (Nowak et al., 2000; Page and Nowak, 2002; André and Baumard, 2011a), fairness in real life is not only characterized by equal divisions. People also consider unequal divisions as fair when they reflect inequalities of skills or talent, or an unequal investment of time, resources, and energy (Schokkaert and Overlaet, 1989; Konow, 2003; Cappelen et al., 2007). Our study offers hints as to why this would be the case. If the reason why humans evolved a sense of fairness is linked to the best way to reward social partners in a biological market (at the ultimate level), each social partner having to be rewarded according to her outside options, then maybe the reason why humans consider that the best contributors should get a bigger part of the benefits is that the best contributors have better outside options in a biological market. An interesting follow-up to our study would thus be to consider the fact that outside options can vary not only because of strength, but also because of skills, talents, effort, etc. Testing this prediction theoretically and empirically would also provide a good entry point to study fairness outside the ultimatum game and its associated always-equal divisions.

Chapter 7

The search for cross-cultural regularities in human fairness (Paper 5)

"50/50 he worked just as hard as I did, and I'm not a jerk."

A4NB9IN89N

7.1 Objectives and summary

The ultimatum game has been widely used to investigate human prosocial behaviors but also widely criticised, notably because it is hard to interpret where the variability in this game is coming from. Because this game is so simple, subjects have to "fill in the blanks" by themselves to interpret what the game is about and a differential filling of the blanks might in part explain the diversity of offers we observe.

As Baumard and Sperber (2010) put it,

> *For example, Henrich et al. (2005)'s study in 15 small-scale societies reveals a striking difference between the Lamalera, who make very generous offers in the Ultimatum Game, and the Tsimane and the Machigenga, who make very low offers in the very same game. But the game is likely to be construed very differently within these societies. The Lamalera, being collective hunters, may indeed see the money as jointly owned by the proposer and the recipient. By contrast, the Tsimane and the Machigenga, who are solitary horticulturalists, may see the money as their own property and therefore feel entitled to keep it.*
>
> [...]

When confronted with cultural differences in experimental result, we should therefore ask: Are they the product of deep differences in the psychological dispositions and processes these experiments are intended to illuminate, or do they reflect differences in the interpretation of the experimental situation? One way to help answer this question would be, for instance, to present the Lamalera and the Machigenga with, as much as possible, the same rich context (e.g., clarifying the source of the money and the relationships between the participants) and assess whether they use the parameters at stake (i.e., rights, past contributions, social links) in the same way.

The goal of this part of my Phd was to grant Baumard and Sperber's wishes: to build an experiment that would be structurally the equivalent of a dictator game (i.e., someone having to unilaterally share an amount of money) but would leave very little room to interpretation, even in different cultures. I chose a dictator game rather than an ultimatum game because the ultimatum game is known to have less variability than the dictator game. But the goal was an ambitious one: is it possible to find a distributive situation in which almost everybody, no matter their culture, would find it fair to share equally?

It turned out that creating a video game was the best way we found to solve many problems of the classical dictator game (see Table 4 P. 116 for an overview of these problems, and http://stephanedebove.net/?p=213 for a playable demo of the game). Notably, the subjects did not know they were taking part in a scientific experiment as they were recruited to "test a video game and give feedback on it". We predicted that in the "dictator video game", almost everybody would offer 50% of the money to be shared. This prediction was not confirmed at all using an mTurk sample: the distribution of offers was exactly the same when people played a classical dictator game or when they played our collaborative video game (in particular, around 30% of subjects gave nothing at all). My interpretation of these results is that many people are on mTurk only to make as much money as possible, and so they do not take into account the fact that they have collaborated to share the money. This interpretation should be tested though, as I have no evidence so far to support it.

The quotations at the beginning of each chapter come from this experiment: they are justifications of subjects who explain why they shared the way they did in our video games (they are not representative of the sample in any way though).

The rest of this chapter comes from a paper in preparation. The core of the video game was not created by me but by Olivier Allouard, a student in video

game design. I did a lot of tweaking to make the game fit for scientific experiments afterwards, and then ran the experiment myself. In the mid to long-term, I also hope the game can be re-used by other researchers to test subjects interacting in collaborative situations.

7.2 Introduction

The dictator game is a reference paradigm in behavioral economics to study humans' prosocial preferences. The game is very simple: it consists in giving an amount of money to one person (the "Dictator") and telling her she has to decide how to share the money with another person she has been randomly and anonymously paired with (the "receiver"). Since the very beginning (Kahneman et al., 1986c; Forsythe and Horowitz, 1994), this game has consistently shown that humans are not entirely selfish and that a substantial fraction of subjects agrees to give something to the person they are paired with (Camerer, 2003).

Because this game is so simple, it has been used to test the influence of a wide variety of parameters on prosocial preferences: demographics (age, sex, occupation...), culture (Western vs traditional societies), social cues (anonymity, degree of proximity...), etc. Engel (2011) provides a useful meta-analysis of 20 years of dictator games. But because this game is so simple, it has also received considerable criticism as to whether its results can really inform us about humans' prosocial preferences. Indeed, the simpler a game is, the less information subjects have as to how to interpret it. As a result, subjects have to "fill in the blanks" by themselves to interpret it and a differential filling of the blanks might in part explain the diversity of offers we observe in the dictator game. In the rest of the introduction, I review the most common complaints made to the dictator game, to show how diverse and subtle pragmatic problems can be (the reader in a hurry might consult Table 4 that summarizes those problems). In the next section, I show how these problems led me to the creation of a video game. I then present and discuss the results of a test of this video game with mTurk participants.

Maybe the most common concern is the use of windfall money, which is far from being an ecological situation (Konow, 2000; Cherry et al., 2002; Frohlich et al., 2004; Cappelen et al., 2007). A dictator who receives windfall money might feel entitled to keep most of it precisely because the experimenter handled it to her. On the contrary, she might feel that she does not deserve it because she had to do nothing to get it. In any case, dictators' divisions will not reflect their "prosocial preferences" but the way they have interpreted the game. Hence, many authors have introduced

a production phase before the dictator game to give more information as to where the money is coming from (Cherry et al., 2002; Frohlich et al., 2004; Cappelen et al., 2007; Oxoby and Spraggon, 2008; Cappelen et al., 2013). A related problem is the problem of contribution assessment. Even if dictators know the recipient played a role in the production of the money, they need to be able to assess how much effort she invested in order to know how to reward her. This is an information that should also be provided to dictators. Finally, even in experiments with a production phase and an assessment of contribution, there is usually no good justification for why only one of the players receives all the money at the end to share it. Ideally, there should be a credible reason for why it is the case, so that the dictator can not believe she deserves the money more than her partner.

Anonymity is another big concern of experimentalists. Lack of anonymity was first conceived as a problem (Hoffman et al., 1994), as it could make dictators generous only for reputational concerns rather than concerns for fairness (see our discussion on this point section 9.1). Hence many dictator games have been done in double-blind settings, where neither the experimenter nor the recipient knows what offers dictators have made. But this double-blind setting has led researchers to suspect a new problem: that dictators might be less generous because they lack information on who their recipient is, a "good guy" or not, "needy" or not, etc. (Eckel and Grossman, 1996; Levitt and List, 2007). Bardsley (2008) reminds us that "dictator game giving provides no evidence of context-free pro-social behaviour or, therefore, orthodox social preferences". An extreme version of this idea is to postulate that dictators keep the money not because they are selfish but because they do not truly believe there is a real recipient who will receive their offer (Frohlich et al., 2001, 2004; Summerville and Chartier, 2012). To avoid "wasting money", dictators who doubt the existence of the recipient could keep all of the money. This concern might be especially important in double-blind experiments where great care is taken not to put dictators in contact with any other subject (subjects are in separate rooms, experimenters use envelopes to transfer money, multiple experimenters are used to interact with subjects).

All behavioral experiments - and not only dictator games - might suffer from what is called the "Hawthorne Effect", "some behavior change as a result of awareness of being a subject in an experiment" (Adair, 1984). The Hawthorne effect might refer to different realities depending on the authors (Adair, 1984), but we can generally identify two effects: "experimenter demands" and "demands characteristics". Experimenter demands refer to subject trying to fulfill the experimenter's expectations. As Orne (1962) puts it,

> *For the volunteer subject to feel that he has made a useful contribution, it is necessary for him to assume that the experimenter is competent and that he himself is a "good subject."*

.

Demand characteristics on the other hand refer to subtle cues which convey an experimental hypothesis to the subject (Orne, 1962). The action set available to the subject could be one such cue in the dictator game. As Bardsley (2008) notes, "subjects might feel that the dictator game is about giving, since they can either do nothing or give". Actually, subjects don't even have to infer what the experiment is about to be more generous than what they would normally be: they could just want to try out the possibilities of action offered to them. Since the subject has come to the lab, and since the only permitted action is to give, why not try to give something?

If both experimenter demands and demand characteristics could affect the results of the dictator game (Orne, 1962; Levitt and List, 2007; Zizzo, 2011; List, 2007; Bardsley, 2008), recently some papers have explicitly addressed the question of the external validity of the dictator game using "natural-field" experiments. Mixed results have been obtained: while some studies show that cooperative behavior in the lab is a good predictor of cooperative behavior in the field (Franzen and Pointner, 2012; Stoop, 2013), other papers report no positive offers at all when people do not suspect they are part of an experiment (Winking and Mizer, 2013). Similarly, "framing effects" depend on what authors call "framing". If framing is about presenting the game as a donation to a charity, framing definitely has significant effects (Engel, 2011). Brañas Garza (2007) for instance reports that adding the sentence "Note that he relies on you" at the end of the instructions increases generosity. More subtle framing relying only on the wording, like calling the game a "giving game" or a "taking game", or calling the action "transferring money", "giving money" or "keeping money", does not change the average offer (Dreber et al., 2013).

A final concern is that subjects could interpret the experiment as a "game", and keep more money just because the goal of many games is to earn as much money, or points, as possible. There is not much we can do about that if we want to keep studying human behavior in the lab, as there is always a risk that subjects do not take the "fake" lab settings seriously. Outside the lab, "natural-field" experiments such as the lost letter paradigm, misdirected letter paradigm (Howitt and Mccabe, 1978; Franzen and Pointner, 2012) or donations to charities (Carlsson et al., 2013) could be used with great benefits.

	Problem	How our paradigm helps dealing with this problem
Konow (2000); Cherry et al. (2002); Frohlich et al. (2004); Cappelen et al. (2007); Oxoby and Spraggon (2008); Baumard and Sperber (2010)	Money is windfall money	Dictators and recipient have to collaborate to produce the money
Frohlich et al. (2004); Cappelen et al. (2007); Oxoby and Spraggon (2008); Baumard and Sperber (2010); Cappelen et al. (2013)	The dictator has no way to assess the recipient's contribution	Contribution is assessable through the avatar's behavior
Frohlich et al. (2001, 2004); Summerville and Chartier (2012)	Dictators may doubt the existence of recipients	Dictators see the recipient's avatar moving
Smith (2010)	The dictator might feel more entitled to the money because the experimenter handled it to her	Who gets to share the money is an unpredictable consequence of the gameplay
Orne (1962); List (2007); Levitt and List (2007); Bardsley (2008)	"Demand Characteristics": Dictators share only because they want to play with the action set available	Dictators do not know they are part of an experiment
Orne (1962); Levitt and List (2007); Zizzo (2011)	"Experimenters Demand": Dictators might try to fulfill the experimenter's expectations and be a "good subject"	Dictators do not know they are part of an experiment
Hoffman et al. (1994)	Lack of anonymity might make dictators generous for reputational concerns	Anonymous experiment over the internet
Eckel and Grossman (1996); Levitt and List (2007)	Lack of social context: double-blind experiments do not give enough information about the recipient	
Frohlich et al. (2001)	Dictators may view the experiment as a game	

Table 4: Criticisms made to the Dictator Game and how our paradigm helps dealing with them

7.3 Methods

We created two cooperative video games, referred hereafter as the "space game" and the "jump game", as well as a replication of the classical dictator game. Games allow two players to play together at the same time, or one player to play with an artificial intelligence. A live demo of the games is accessible online at

http://stephanedebove.net/?p=213. The video games were coded in HTML5 (javascript), and are released under MIT license. Source code is fully available online at https://github.com/BigNoob/DEMOG. The main technical challenge when building real-time video games is to keep low latencies, to ensure a smooth and consistent user-experience. This was made possible thanks to the development of relatively recent technologies such as Node.js and Socket.io.

7.3.1 Description of the video games

In the space game, each subject controls a spaceship located at the bottom of the screen, and can move it left and right (see screenshot on Fig. 14A). The goal of the game is to kill all enemy spaceships located at the top of the screen by moving with the arrow keys and pressing the space key to shoot. When an enemy spaceship is killed by either one player, the common score of the two players is increased on the mothership. The mothership can only be killed after every other ship, and will fall to the ground when hit. The player closer to the fallen mothership is able to grab the points and access a sharing screen where she is asked to indicate how she wants to share the points.

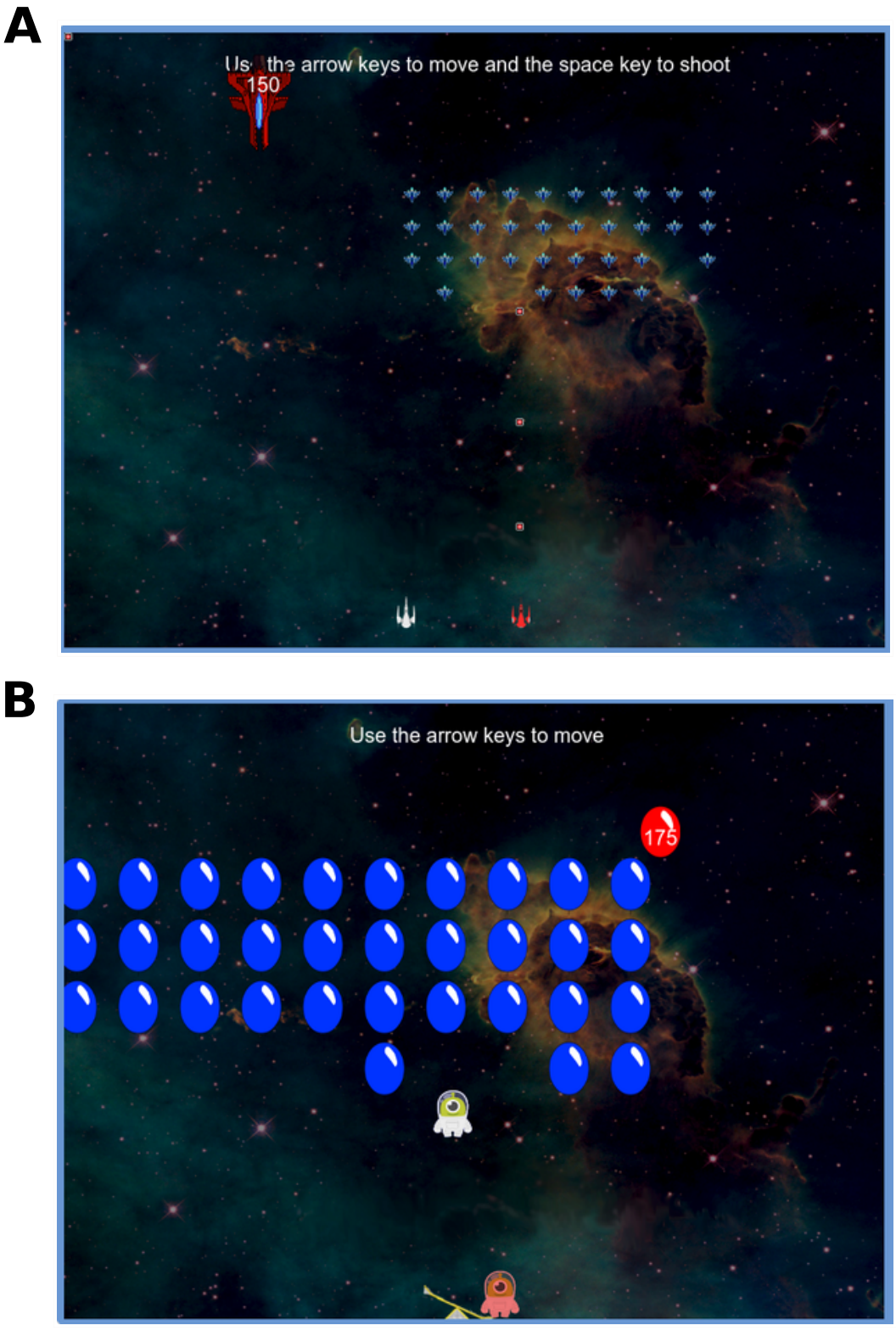

Figure 14: Screenshots of the two video games. A: the "space game" in which the goal is to kill all enemy spaceships (players control the ships at the bottom of the screen). B. the "jump game" in which the goal is to pop all of the balloons (players control the seesaw at the bottom of the screen)

In the jump game, each subject controls a character located at the bottom of the screen, while balloons are flying in the air (see screenshot on Fig. 14B). The goal of the game is to jump on the seesaw to pop all the balloons. Each player has to move in turn the seesaw by pressing the left and right arrow keys so that the other player lands on it. Each time a balloon is popped, the common score of the two players is increased on a master balloon. The master balloon can only be popped after every other balloon. The player who catches the master balloon at the end is able to grab the points and access a sharing screen where she is asked to indicate how she wants to share them.

Additionally, to serve as a baseline condition we created a replication of the classical dictator game by having a "game" consisting of only the sharing screen.

7.3.2 Generic advantages of video games

The main advantage of video games is to avoid that subjects suspect they are taking part in a scientific experiment, while at the same time being able to control precisely the environment they are interacting in. In this sense, video games combine the advantages of lab experiments (careful design, randomized treatments) with the advantages of "natural-field" experiments (no demand characteristics or experimenter demands). On the internet, experimenters can present their video games as needing beta-testers and feedback to improve them. In the lab subjects necessarily know they are here for science, but it is still possible to hide the fact that the experiment is about *sharing* by making the sharing part of the game only a small part of the entire game. Video games also allow to build environments that are less poor in stimuli than traditional environments, and to record an unlimited number of variables regarding the subject's behavior (such as, in our space game, the number of shots fired, the number of enemy killed, the distance covered, etc). Finally, video games developed in HTML5 language can be played in any browser, and thus on many platforms, including tablets and smartphones.

7.3.3 Specific advantages of our video games

How our specific design helps dealing with the problems of interpretation mentioned in the introduction is summarized in Table 4. First, the money is not windfall anymore but players have to collaborate to produce it. Note that our design presents an important difference with previous dictator game experiments involving a production phase: in our experiment, players have to *collaborate* on the same task and their effort increments a *common* gain. To our knowledge, all previous studies had subjects work relatively independently to produce the gains (correcting spelling

mistakes, answering quizzes) and had the gains of the recipient *transferred* to the dictator afterwards. Nonetheless, players can still work relatively independently in our space game (they do not need each other to kill enemy spaceships). This is why the jump game exists. With this game, we wanted to capture the ecological situation in which two people necessarily need each other to produce a benefit (like climbing on someone's shoulders to reach a honeypot). In the jump game, one player alone can not pop all the balloons.

In our games, dictators have less reasons to doubt about the existence of recipients as they see the recipient's avatar moving on the screen. Why only one player gets the whole money at the end is now credible: the common points are stored in the mothership (for the space game) or in the master balloon (for the jump game). Who gets to share the money is not decided by the experimenter but is a natural consequence of the game design. In the space game, the mothership falls in the direction it was moving just before being shot. In the jump game, it is impossible to predict who will pop the last balloon, and thus who will get the points. Finally, the contribution of each player can easily be assessed by displaying the number of enemies killed by each player (although we will not study this parameter in the present paper).

7.3.4 Experimental procedure

We recruited 131 subjects on Mechanical Turk (mean age 34, standard deviation 11, 82 males and 49 females). Each subject played only one game. 42 subjects played the classical dictator game, 45 played the space game, and 44 played the jump game. Subjects tested on the video games did not know they were taking part in a scientific experiment, they were recruited to "test a video game and give feedback on it". Subjects tested on the classical dictator game were recruited to "answer a survey". Upon arrival to our website, subjects learn that the game is a two player game and that they will have to wait at the beginning of the game until another Mturker connects. Subjects learn that they will be paid a flat fee of $0.5 for testing the game, plus a bonus for all the points they earn in the game: 1000 points are worth $1. Using such small stakes has not been shown to increase offers (Amir et al., 2012) or only very slightly (Novakova and Flegr, 2013; Raihani et al., 2013).

Subjects then read short instructions (illustrated with screenshots) about how to play the game, and enter a waiting room. A game starts as soon as two players are in the waiting room. If a player waits for more than 90 seconds, she is automatically removed from the waiting room and starts a game with the artificial intelligence. The subject is not informed she is playing with an artificial intelligence, and it is hard to

detect it just based on the artificial intelligence's avatar movements, but we decided a priori to exclude those subjects from our analysis. To prevent subjects from playing the game multiple times (either because they liked it or because they were angry at the offer they received), we recorded subjects' IP addresses and prevented subjects to enter the waiting room if their IP had been seen before. This also prevented subjects to try to play the game with themselves (i.e. by opening a new tab or window). We also checked that our data did not contain any duplicate Mturk ID. A game could not start if one of the two partners was not keeping her game's tab active (possibly because she was multitasking and browsing in another tab). This allowed the two players to be active in the game from the very first seconds. We sampled every 250 ms what keys each player was pressing, to make sure the players remained active.

After the game, subjects were redirected to a questionnaire where they were asked about their age, sex, and a justification about why they shared the points the way they did. In the video game conditions, subjects were also asked to rate on a scale from 1 to 10 how hard and how fun they found the game, and if they found any bugs in it. In the classical dictator game condition, they were also asked, following Rand et al. (2012), whether they had already participated in similar experiments before. Those were exploratory variables as we had no a priori hypotheses on their impact. Post-game feedback does not give any reason to think subjects suspected they were part of an experiment.

7.4 Results

The results do NOT confirm our predictions (at all). Figure 15 shows the distributions of offers made by dictators in our three conditions. Distributions look the same, and a Kruskal-Wallis test fails to reject the null hypothesis that the samples originate from the same distribution ($K = 0.61$, $p = 0.74$). Pairwise comparisons using Mann-Whitney tests also fail to reject the null hypothesis, showing no significant difference between the medians of the classical dictator game and space game conditions ($U = 860$, $p = 0.45$), classical dictator game and jump game conditions ($U = 902$, $p = 0.84$), space game and jump game conditions ($U = 929$, $p = 0.60$). Our prediction that the variance of the distributions could be reduced in the video games compared to the classical dictator game is also not confirmed by statistical tests. Conover tests fail to reject the null hypothesis that the variances of samples taken two by two are equal (classical dictator game vs space game, $C = 0.72$, $p = 0.47$; classical dictator game vs jump game, $C = -0.45$, $p = 0.65$; space game vs

jump game, C = 0.51, p = 0.61).

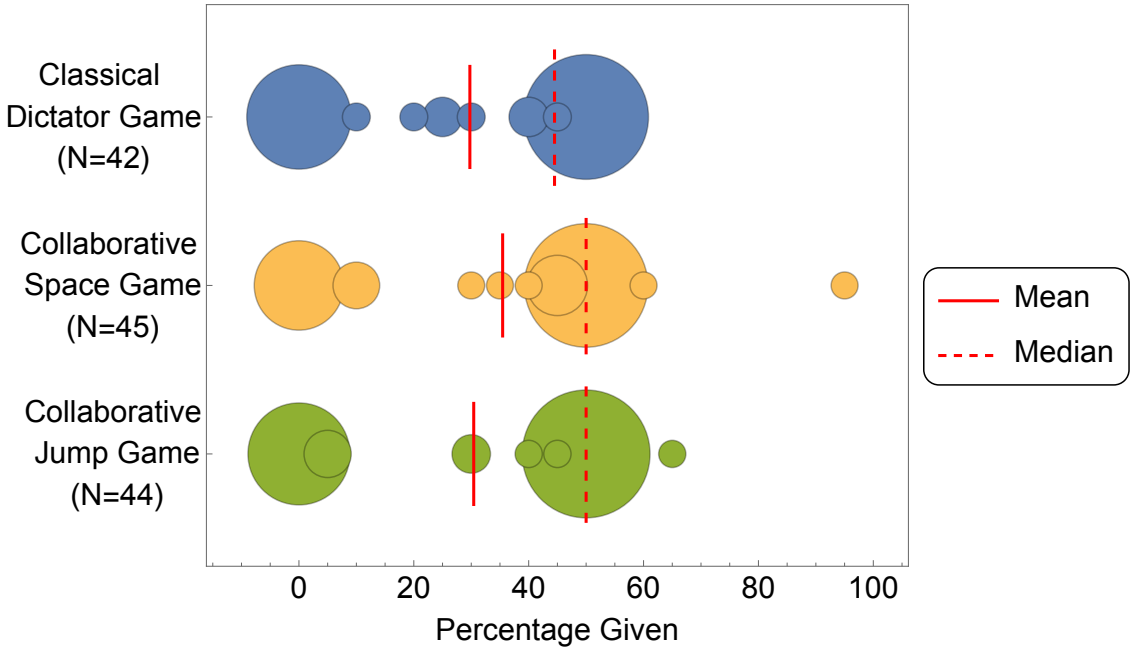

Figure 15: Distribution of the offers made by the dictators in our three conditions. The area of the bubbles represents the percentage of such offers made by dictators. For instance, the largest blue bubble represents 48% of the total number of offers made (i.e., 48% of dictators offered half of the points in the classical dictator game condition).

7.5 Discussion

Like most theoreticians, I don't like when the data does not confirm the predictions of the theory. It was rather surprising to me that collaborating more than 2 minutes before sharing the money had no impact at all on the amount given. My first reaction was to look at the justifications "selfish" people gave. In my experience, justifications for why subjects share the way they share are rarely asked by economists (or at least rarely reported). There could be some good reasons for this, as it has been debated for decades whether or not introspection could be used as a scientific tool in psychology (Hatfield, 2005; Sackur, 2009). But it might also be because of economists' focus on payoffs and strategies, with no regard for the mental processes that led to the production of such strategies. In any case, introspection can help us generate new hypotheses that can be tested with less controversial methods afterwards.

Figure 16 presents the frequencies of different categories of explanations, for

dictators who offered 20% or less. The 20% threshold is arbitrary but at the same time corresponds to the often cited threshold under which offers have a very large probability to be rejected in the ultimatum game. Note that sample size is small (between 14 and 17 subjects depending on the condition) and coding was made manually by me only so the results have to be taken with caution, it is a very exploratory analysis. A variety of explanations is given as to why dictators behaved selfishly, but the most prominent one, accounting for around 50% of the results in each condition, is just that the dictator admitted to be "selfish", "greedy", to want to "maximise her bonus" or to "keep all of the points" for herself.

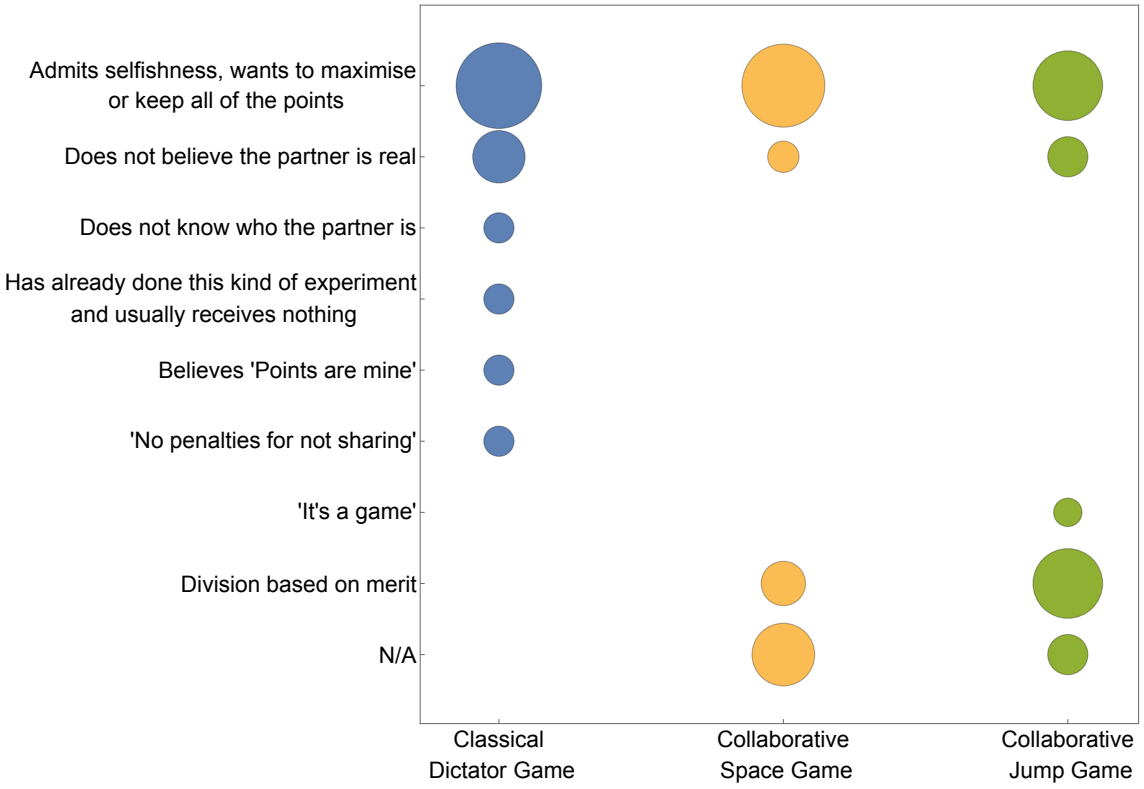

Figure 16: Frequency of justifications for dictators offering less than 20% (in % of total frequency).

I think this open admission of selfishness and desire to maximise one's gains is enlightening. It reminds us that the Mturk community, although more diverse than the university students (Ross et al., 2010; Ipeirotis, 2010), is still constituted to a large extent by people who are here "only for the money". In a survey of mTurk users, Ipeirotis (2010) reports that mTurk is a primary source of income for 13% of mTurkers and a secondary source of income for 61% of mTurkers. Ross et al. (2010) show that the percentage of people who declare either that "MTurk money is

always necessary to make basic ends meet", "MTurk money is sometimes necessary to make basic ends meet", or "MTurk money is a way for me to pay for nice extras" is 38% (38% is also, incidentally, the percentage of people who gave nothing at all in my experiment). While this is no ultimate evidence at all, it makes sense to think that if some people are on mTurk only to make money, collaboration will have little impact on the way they share.

As I explained in the introduction of this chapter, our original goal was to design a dictator game to reduce the cross-cultural variance of offers. Because we thought that a collaborative video game would be the best way to deal with the pragmatics problems of the dictator game, we progressively drifted towards testing internet samples, and in particular mTurk users. Unfortunately, by doing so, we introduced new problems by biasing our sample. Hence, one lesson this experiment teaches us is one that is familiar to many behavioral scientists: avoid testing WEIRD samples (Western Educated Industrial Rich Democratic) or at least know exactly what your sample is made of (Henrich et al., 2010).

This experiment also made me realize that although the results of the dictator game are often described as "40% of people give nothing", this is only true in Western student populations. In a meta-analysis of 131 dictator games, Engel (2011) plots two very interesting graphs (reproduced in Fig. 17), comparing offers made by students and non-students, and comparing offers in different classes of age. The results are striking: it is mostly students who make null offers (or people of student age). Age is the factor that has the most impact on the offers in the whole meta-analysis. In particular, middle aged people's offers are remarkably clustered around the fair offer of 50%. 40% of old people also give *everything* to the recipient, while virtually none of them gives something smaller than 50%. Hence, the picture of human fairness might be less grim than what is already thought. Although I have nothing to back this up, I suspect that the 40% of students who give nothing in lab experiments are also, in large part, only doing the lab experiments for the money (and it would be interesting to compare the offers of students who are paid in money versus students who are paid in university credits). I have myself plotted the offers as a function of age in my experiment (Fig. 18), but the statistical analysis on this point remains to be done.

The last lesson I will keep from this experiment is that we should pay more attention to the *distributions* of offers. Many papers rely only on means and standard deviations to describe their results. Yet, it is particularly striking in my results (Fig. 15) and to some extent in the meta-analysis (Fig. 17) that the distributions have modes. In my results, around half of the participants offer 50% and 35% offer

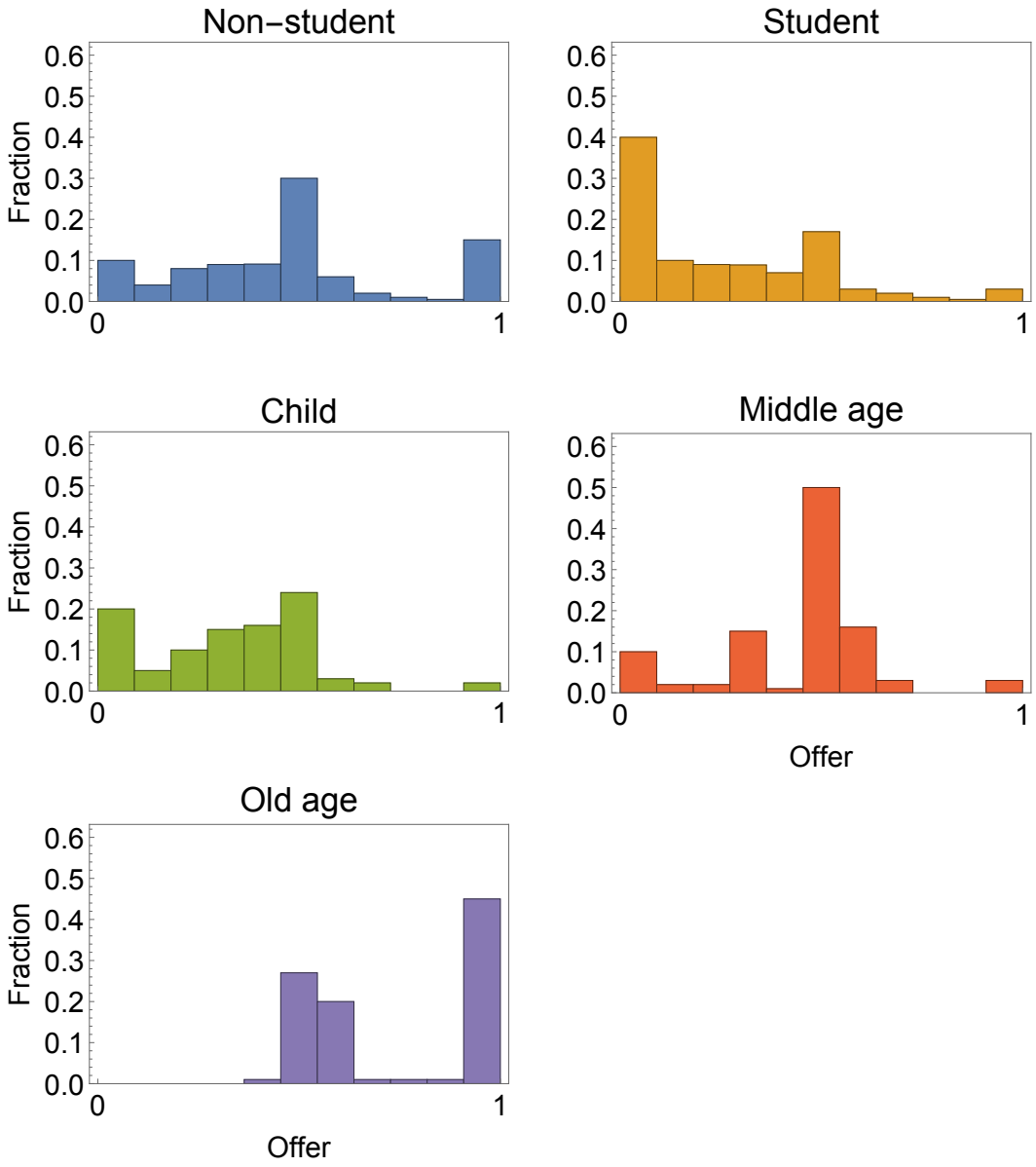

Figure 17: Distribution of offers in the dictator game for different demographies of population. Data extracted visually from the figures of Engel (2011), so the bar heights are only approximative.

nothing, which leaves little room for intermediate offers. This could be an argument for the existence of a sense of fairness, as opposed to the existence of a continuum of ways to cooperate between over-selfishness and over-generosity. Said differently, there is little variation in the offer that people call fair. Looking at the justifications, people who give something between 20% and 45% are people who wanted to be selfish but felt guilty at being so, and wanted to give "a little something" to their recipient. But those people did not call their offer a fair offer.

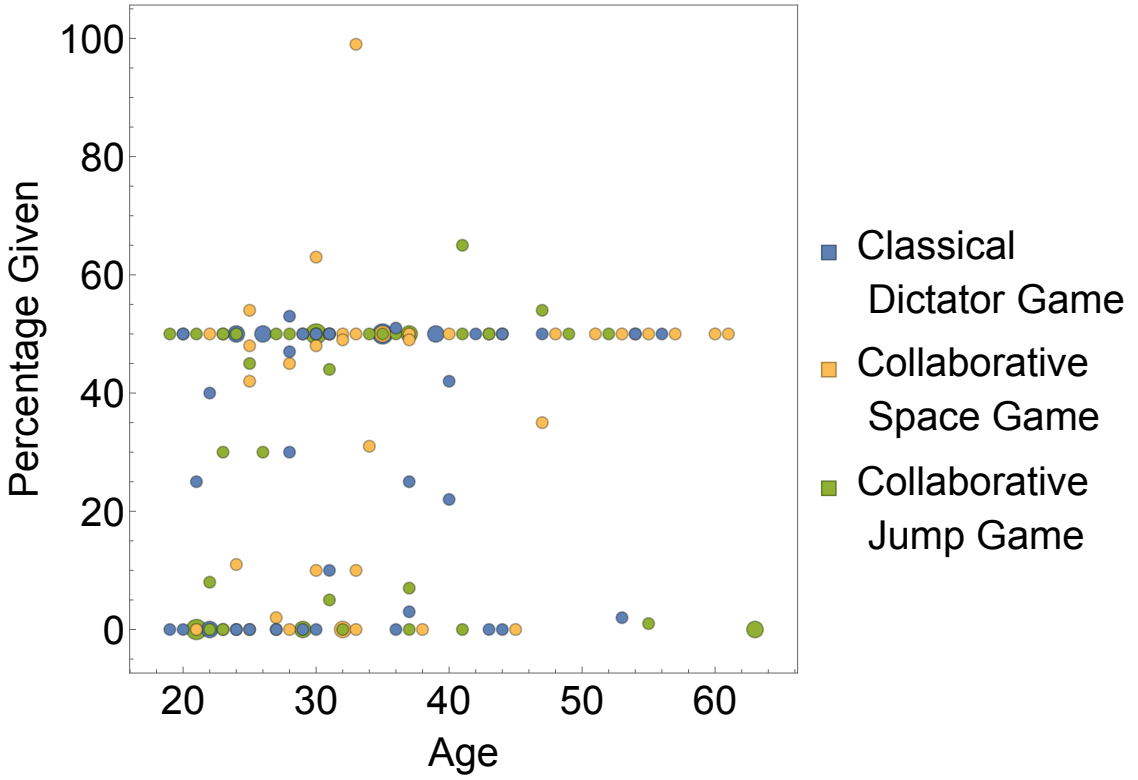

Figure 18: Distribution of the offers made by the dictators in our three conditions, as a function of age. The area of the bubbles represents the percentage of such offers made by dictators.

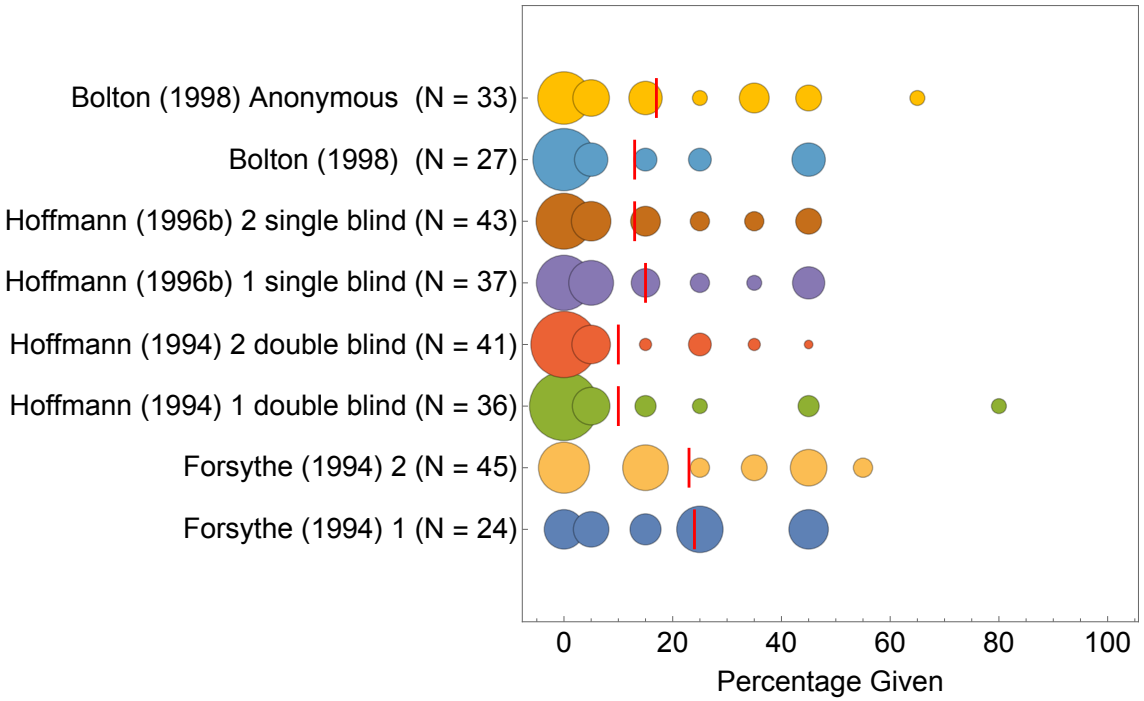

Figure 19: Examples of distributions found in the first historical dictator games.

Nonetheless, it is unclear how widespread those modes really are outside of my experiment. Although modes do appear in the meta-analysis by Engel (2011) (Fig. 17), there is considerably more variation in this meta-analysis than in my experiment, especially in the student population. For informative purposes, I have also plotted the offers made in some of the first historical dictator games (Fig. 19). They do not seem to present the "fair" mode at 50%. Caution should thus be taken as to how to interpret the distributions, but studying distributions is something that deserves more attention in my opinion.

From reading the feedbacks, it looks like I was not completely successful at eradicating the pragmatic problems I was talking about in the introduction. For instance, even in video games where subjects saw the other player's avatar moving, two subjects doubted they were really interacting with someone real (but 2 subjects out of 131 only). Nonetheless, this could not predict how fair they would be, as one subject justified her offer by saying *"I wasn't sure there was another person, so I didn't want to just give my potential bonus away."* and the other by saying *"I shared because I wasn't sure if the other player was real or not"*. Note also that subjects really are habituated to economic games on mTurk, and that this could influence the results. When asked to justify her offer, one subject said *"I didn't have to share points, but if I did I would've shared half if that's all I had to do, and if it was*

multiplicative with how much I could share to them and they could share back then I would have shared all of them and trusted them to split it evenly with me.". In other words, this subject was giving her hypothetical behavior in a *trust* game even if I did not test her on a trust game...

The dictator video game is one of the projects I have done last in my PhD, so the analysis is not finished and I do not have any evidence to support the interpretations of the results I have made above. But the fact that the results did not follow my predictions has motivated me to dig deeper in the literature and generated new ideas that I hope I will be able to explore in the future.

Part V

General discussion

Chapter 8

If partner choice is not limited to humans, why would fairness be?

"I did not share. If I was forced, I would give the minimum amount. It is best to take care of yourself first."

A892NJV24NRS2

There is no simpler way to present the topic of this chapter than Bshary and Raihani (2013)'s comment on Baumard et al. (2013)'s paper :

> *Baumard et al. propose a functional explanation for the evolution of a sense of fairness in humans: Fairness preferences are advantageous in an environment where individuals are in strong competition to be chosen for social interactions. Such conditions also exist in nonhuman animals. Therefore, it remains unclear why fairness (equated with morality) appears to be properly present only in humans.*

It is of course a big question to tackle, and I prefer to acknowledge right away that this enterprise is far from being finished at the time of writing this thesis. I will only scratch the surface of the literature here. In the first section I review the evidence for the existence of a sense of fairness in non-human animals, which leads me to conclude that there is not much evidence. In the second section I summarize some of the hypotheses that have been suggested to explain the absence of a sense of fairness in non-human animals.

8.1 The (non-)evidence for fairness in non-human animals

8.1.1 Fairness in the lab

Seeing in Chapter 7 how difficult it was to interpret economic games in humans, I am not much optimistic about what they can tell us about fairness in non-human animals. Can we decide beforehand what result would count as a definite proof of the existence of a sense of fairness in non-human animals? Negative reactions to disadvantageous inequity can come either from fairness or envy (Christen and Glock, 2012), and envy will remain the most parsimonious explanations. Negative reactions to *advantageous* inequity would be more interesting, but without access to the motivations of the animals, we will never be sure that they result from fairness rather than the fear of retaliation from the disadvantaged party. Nonetheless, laboratory experiments are the best tools we have to compare behaviors across species at this time. The basic idea behind all experiments is to have a subject and its partner invest different efforts and/or receive different rewards, and observe if and how the subject will react to combinations of effort/reward that humans would find unfair (see Brosnan and de Waal (2012); Bräuer and Hanus (2012) for a review in primates).

The ultimatum game

The ultimatum game has been tested on different non-human primate species. The results have been mixed. An early report by Jensen et al. (2007) found that chimpanzees behave as rational-maximisers in the ultimatum game. One proposer chimpanzee had to choose between pulling two trays: one with a fair equal division of food reward (50/50) and one with an unfair but favorable division (80% for the proposer). Proposers almost always chose the most advantageous division for themselves, and responders almost always accepted this division. Proctor et al. (2012) modified the experiment by having chimpanzees learn that one token could be exchanged for a 3/3 distribution of food items and another token exchanged for a 5/1 distribution in favor of the proposer. To get the reward, the proposer had to select a token and pass it to the responder AND the responder had to pass the token back to the experimenter. In this situation, the proposers chose the fair token 75% of the time. Nonetheless, in a second condition in which the responder was present but passive (the proposer could pass the selected token directly to the experimenter to get the reward), the proposer chose the unfair token almost 90% of the time, which suggests that proposers were not that fair after all. To summarize, in both Jensen

et al. (2007) and Proctor et al. (2012) responders always accepted any offer, but only in Proctor et al. (2012) did proposers make fair offers. To explain this difference, Proctor et al. (2012) suggest that their paradigm involving tokens rather than trays is more intuitive to understand, but Amici et al. (2014) have recently shown that primates do not really understand this task.

Inequity experiments

Inequity experiments involve giving less to a primate than what his partner receives and observing his reaction (Brosnan and De Waal, 2003; Brosnan et al., 2005). There are two important results in this literature. When primates do not have to work to get their rewards, they do not show any concern for disadvantageous inequity (Bräuer et al., 2006; Dindo and De Waal, 2007; Bräuer and Hanus, 2012). Rare cases of frustration seem to be caused by individual expectations about what one could have received given the food present in the room, rather than social concerns. In fact, some primates being inequitably treated responded by an increase of consumption of the less-preferred food. On the other hand, when primates do have to work to get their rewards, negative reactions to disadvantageous inequity have been observed in chimpanzees, bonobos, capuchin monkeys, macaques, but the results have been mixed and inconsistent, even within the same species (see Brosnan (2013); Bräuer and Hanus (2012) for a full list of references). Some have even argued that group-specific traditions could explain this inconsistency, so great care should be taken as to how to interpret those results.

If those results were confirmed though, Brosnan (2013) notes that negative reactions to inequity in different species seem to correlate positively with cooperation among non-kin: chimpanzees, bonobos, capuchins and macaques responded negatively to inequity, and they all cooperate with nonkin, at least for coalitions and alliances (Brosnan, 2013). On the contrary, squirrel monkeys and orangutans, which lack cooperation with nonkins, or night monkeys and tamarins, which cooperate but mostly through biparental care, do not show an aversion to inequity[1]. Finally, if an effort seems to be important to trigger negative reactions to inequity, how hard a primate has to work to get his reward does not impact his refusal rate. When the subject has to work to get food while the partner is given food for free, no negative reactions are recorded. Hence, if effort is taken into account, it seems to be taken into account qualitatively but not quantitatively.

[1] which would also suggest this behavior is not a homology within the primates.

Prosocial games

In the previous games, subjects had no possibility to restore fairness even if they wanted to. In prosocial games they can: the subject decides between a 1/1 division, a 1/0 division in her favour, and a 1/2 division in favour of her partner. Prosocial games are often presented as the primate equivalent of dictator games, but they can hardly be so as they involve giving food at no cost to oneself. Once again, results are mixed, even within the same species. In barpull tasks (where the rewards are on trays that have to be pulled), a robust finding is that chimps do not preferentially choose the 1/1 option over the 1/0 option: they do not provide more food to others even if this does not affect their reward (Silk et al., 2005; Jensen et al., 2006). Nonetheless, they do prefer the 1/1 option when the options are materialized by tokens rather than food on trays (Horner et al., 2011). Capuchin monkeys and long-tailed macaques also seem to prefer the 1/1 option, although in the case of capuchin monkeys, this preference is stronger towards the subject's own kin. Chimpanzees and cotton-top tamarins do not distinguish between a 0/0 and 0/1 disadvantageous distribution (Jensen et al., 2006; Stevens, 2010), but marmosets do (Burkart et al., 2007).

An interesting result is that in no study except Brosnan et al. (2010), primates reacted to advantageous inequity. Brosnan (2006) even reports that *"in several situations in which the [disadvantaged] subject rejected the cucumber slice, the [advantaged] partner would finish their grape and then reach through the mesh to take the subject's cucumber and eat it as well!"*.

More recently, in a study comparing 15 different species, Burkart et al. (2014) showed that those prosocial behaviors are best predicted by the extent of allomaternal care in each species.

Summary

As Bräuer and Hanus (2012) put it, we are faced with a "rather fuzzy empirical picture". Maybe I should have started my review with this, but in a recent paper, Amici et al. (2014) controlled meticulously for all methodological flaws that had been reported in previous papers, and tested six primate species on both the barpull and token task. Their conclusion: "Our results provided no compelling evidence of prosociality in a food context in any of the species tested.". They report for instance that when appropriate controls are added "most of the subjects did not understand the [token] task". I am not an expert of the field, but it seems to me that their study cast doubts on many positive results that I reported above.

From the historical sequence of articles falsifying each other in the last ten years, there seems to be a cleavage between researchers who think we can already attribute fairness concerns to non-human primates and those who think we can't. In fact, authors seem to agree on the results but to draw different conclusions. For instance, Brosnan and de Waal (2012) in a review paper write:

> *"Do non-human species have a sense of fairness? We believe that the answer is a qualified 'yes.' They do show behaviors indicative of both a sensitivity to their own and others' outcomes, however, these responses are neither as consistently elicited as those seen in humans, nor do they appear to be as strong, particularly in the case of second order fairness [(advantageous inequity aversion)]."*

which is similar to what Bräuer and Hanus (2012) say after their own review, but the conclusion is different:

> *"The results [...] provide only weak evidence for a sense of fairness in non-human primates. Although apes and monkeys are attentive to what the partner is getting, they do not seem to be able or motivated to compare their own efforts and outcomes with those of others at a human level."*

My own opinion is that the existence of "behaviors indicative of both a sensitivity to their own and others' outcomes" does not seem a reason good enough to attribute a qualified 'yes'. A sensitivity to one's own and others' outcomes can be useful for other mental mechanisms than a sense of fairness, or produced by other mental mechanisms than a sense of fairness. My own reading of the literature would make me attribute a qualified 'no', qualified because I feel like ten more years of research would not be too much to be able to answer this question.

As evoked in the previous section, prosocial behaviors in primates could be the result of convergent evolution rather than be homologous to each others (Brosnan, 2013). Hence, we do not necessarily want to look for fairness in our closest related species but in the species sharing the same selection pressures. Experimental tests of fairness in non-primate animals are starting to appear, but the data is quite limited so far. If methodological flaws still exist in primate research on fairness after ten years, it is plausible that methodological flaws also exist in other animal studies. For the record, negative reactions to disadvantageous inequity have been reported in domestic dogs in one study (Range et al., 2009, 2012) but not another (Horowitz, 2012). They have also been reported in corvids (Wascher and Bugnyar, 2013), but not in the cleaner fish *Labroides dimidiatus* (Raihani et al., 2012; Raihani and McAuliffe, 2012). If methodological problems can be solved, an exciting field of research seems to be opening up, and I am looking forward to seeing how it develops.

8.1.2 Fairness in the real world

If a sense of fairness is already difficult to identify in the lab, it might be even harder to identify in natural situations. Fair outcomes might be easier to identify than a sense of fairness, but they can usually be explained more parsimoniously than with the existence a sense of fairness. In theory, a good way to start would be to identify negative reactions to unfair situations, or situations in which one could have taken advantage of one's strength but did not. In practice, it is really hard to do.

In most animal species, dominance constrains the distributions of benefits. Dominant individuals get more food, sexual partners, territories... But is this necessarily unfair? Dominant individuals get more but they usually also bring more to the group, by better defending the group in intraspecies conflicts for instance. Hence, we can not rule out the possibility that it is "fair" for them to get most of the benefits. This situation is actually reminiscent of Fiske (1992)'s authority ranking relationships in humans. In authority ranking relationships, people in high rank have a privileged access to resources over those in lower rank, but the privileges also come with duties: usually, the duty to protect the lower ranks. For instance, Shweder et al. (1997) argue that if lord/servant relationships in the feudal systems are seen as unfair today, they were not always perceived this way, because each party had duties toward the other. Lord/servant relationships could be analogous to the primates' dominance systems, with the important difference that there is no reason to think that non-human subordinates feel they have the *duty* to give more resources to the dominants, or that dominants think they have the *right* to take more. This makes all the difference, and once again brings us back to the difference between a fair outcome and a fair psychology.

Similarly, when dominants do share with less dominant individuals, it is not easy to determine whether this behavior could come from a sense of fairness. Reproductive skew models have shown that the dominants' selfishness is constrained by the outside options of subordinates, i.e. their opportunities to join other groups. Because dominant individuals have a direct interest in keeping subordinates in their group (for food foraging, protection, grooming...), their behavior might look fair without being motivated by a sense of fairness at the psychological level. It is also difficult to assess in the wild whether individuals are pregnant or lactating, have recently eaten, mated, or been groomed, all of which can potentially explain fair sharings without implying the existence of a sense of fairness (Brosnan, 2006). The difficulty to measure costs and benefits when dominance relationships exist was already noted by scholars trying to identify what would be good examples of reciprocity in nature (Seyfarth and Cheney, 1988; Silk, 2006).

Brosnan (2013) takes policing behaviors in primates, in which dominant males intervene in fights on the side of the loser, as "some of the best evidence in favor of a sense of fairness", indicating "that these males recognized social inequalities in others' interactions and were willing to act against their own short-term self interest to rectify them."

To me, it is difficult to imagine how these behaviors could be interpreted in favor of the existence of a sense of fairness. Recognizing social inequalities and acting against one's own short-term interest are two behaviors that can be explained by more parsimonious psychological mechanisms. Bshary et al. (2002) remind us that policing behaviors exist in fish too: when cichlid *N. multifasciatus* females try to prevent other females to enter their group, males sometimes intervene in favor of the newcomer and increase the newcomer's probability to enter the group. We have no evidence for the non-existence of a sense of fairness in fish, but the presence of those behaviors in fish reminds us that following "simple rules", such as always defend a newcomer, could be enough to explain these behaviors (Bshary et al., 2002). As Bshary et al. (2002) suggest, we could then use these simple rules as 'null hypotheses' to see whether or not they can also explain primate behavior.

In natural settings, gorillas have been shown to be "sensitive to inequities during their naturally occurring social interactions" (Van Leeuwen et al., 2011). During play fights, one gorilla usually comes close to another one and hits him. The authors observed that the hitter then starts to run *first* to keep playing. They interpret this behavior as showing that gorillas understand they have a "competitive advantage" when they hit first, and they are "trying to maintain it" by running first. On the contrary, they would have interpreted the hit gorilla running first as an "evidence of inequity aversion" (p.39, last paragraph of the introduction). The authors' interpretation in terms of inequity goes too far in my opinion. No scientist would seriously deny that primates can recognize "differences" in general, differences of power, of food quantities, of mate values, and that they can act according to these differences, in a direction that humans would find fair or not. These studies emphasize the necessity to be very clear about what behaviors we should expect to be produced *only* by a sense of fairness.

Finally, many responses to unfair outcomes might come from deviations to an individual's expectations rather than be produced by a sense of fairness. As Bshary and Raihani (2013) put it:

> *"While fairness preferences imply that individuals monitor and respond to the relative payoffs accruing to themselves and to a partner, a simpler alternative is that individuals have an internal expectation about*

payoffs from an interaction and adjust their behaviour (e.g., by switching partners) if these expectations are violated (Chen and Santos, 2006). Crucially, responses based on fairness preferences and responses based on self-referent loss aversion can both lead to cooperative and fair outcomes."

Bshary and Raihani (2013) report that different cleaner fish jolt the same amount in different cleaning stations, as if they had made a contract and were respecting it. But they interpret this "fair" situation as due to "individual learned optimization of own payoffs by both cleaners and clients": cleaners have learned individually the maximum amount of jolting they can do before their client leaves. Cleaner males also punish females when females cheat during a common inspection, but this punitive behavior can still be based psychologically on loss aversion (the client's departure) rather than fairness (the female received more than the male). I suspect many negative reactions we observe in primate experiments can also be explained in terms of disappointed expectations.

8.1.3 Summary

From this small preliminary review of fair behaviors in non-human animals, I see three distinctions that a more extensive review could make to help clarify the literature:

- a distinction between fair outcomes and fair psychology. If the distinction is easy to make, it is in practise extremely difficult to know whether a behavior has been produced by a sense of fairness or not. As I noted above, knowing this often requires to know whether the animal thinks it has the duty to behave in a non-selfish way, so I am not much optimistic about our chances to solve this problem. Maybe future progresses in neuroscience can help address this problem though.

- a distinction between simple rules that are domain-specific and a more general sense of fairness that applies to a large range of situations

- a distinction between behaviors that provide immediate benefits and behaviors that provide benefits via an enhanced reputation. This opens the door for simple rules to be categorized as genuine fairness even though they are not domain-general. For instance, even though male *N. multifasciatus* defends female newcomers following a simple psychological rule, the important question would be to know whether this rule evolved because the males have a stake in the mere presence of the newcomer (cooperation for immediate benefits) or

because it helps to obtain a good reputation and reap benefits that depend on the later *investment* (and choice) of an other individual.

8.2 Hypotheses for the lack of fairness in non-human animals

8.2.1 Lack of cognitive capabilities

Many scholars explain the lack of reciprocity in non-human animals in terms of lack of cognitive capabilities (Dugatkin, 1997; Stevens and Hauser, 2004). Cognitive abilities have been pointed out to explain the lack of fairness too. Bräuer and Hanus (2012) for instance suggest that non-human primates might lack the capability to compare their own outcome or their own effort with the one of others. But explanations relying on cognitive skills can be problematic. First, many supposedly complex cognitive skills are found in non-primates. For instance, one could think that remembering the cooperative behaviors of different partners at the same time is cognitively demanding - yet these capabilities have been observed in fish (Dugatkin and Wilson, 1993; Dugatkin, 1997; Bshary et al., 2002). Second, from an evolutionary perspective, some cognitive limits are more interesting than others. If there are no good reasons as to why some cognitive mechanisms might not have evolved, postulating that fairness did not evolve for cognitive reasons only displaces the question - or just solves it at the proximate level (André, 2014). We can still ask why those cognitive abilities did not evolve outside humans. The low-level explanation that it was a matter of mutational chance can never be completely ruled out (Ridley, 2004), but it might also be that the required selection pressures were only present in humans.

Christen and Glock (2012) and Sperber and Baumard (2012) have suggested that a theory of mind, whose existence in chimpanzees is only partial (Call and Tomasello, 2008), might be one missing cognitive skill:

> *"Predicting an individual's future behaviour just on the basis of her past behaviour would ignore psychological factors that, in the human case, are crucial. A mere behavioural assessment may be good enough in other animals' repetitive forms of mutualistic cooperation (as between cleaner fish Labroides dimidiatus with client reef fish—see Bshary and Schäffer, 2002). In the human case however, given the open-ended variety of forms and conditions of cooperation and the complexity of people's beliefs and*

motivations, cooperativeness cannot be effectively assessed without making inferences about others' mental states and dispositions" (Sperber and Baumard, 2012)

Nonetheless, it seems to me that a sense of fairness that would only be based on behaviors (effort and outcomes), but not intentions and beliefs, would already be beneficial, and thus could be present in primates.

Another good candidate that we are sure is missing in non-human animals is language (Melis and Semmann, 2010). Language is used to exchange information about others and is necessary to build reputations, if "reputation" is understood not in the narrow sense of the opinion that I have of someone else, but in the broader sense of the "socially transmitted [...] judgment [about someone] that is presented as consensual, or at least as widely shared" (Sperber and Baumard, 2012). We do not model the formation of reputation in our models (see section. 9.1 for a discussion) but it is plausible that this kind of reputation is highly important for the evolution of fairness. Without language to communicate the unfair actions of others, unfair behaviors might pay off. Without language, it might also pay off to be fair in a Machiavellian way rather than in a genuine way, because the cost of making mistakes (to be unfair when one was observed) will not necessarily be high. The fact that up to 38% of our conversations is devoted to personal relationships lends support to this idea (Dunbar, 1993).

8.2.2 Other hypotheses

I list here a few hypotheses for the non-existence of a sense of fairness outside humans. They come either from the literature or from our models (because we found a critical parameter in our models for fairness to be able to evolve). This list is of course non-exhaustive.

- from Baumard et al. (2013): *"the much narrower and relatively fixed range of mutually beneficial interactions occurring in non-human species does not result in the social selection of a general and hence properly moral sense of fairness"*.

  Insisting on the varied forms of cooperation in humans also makes the problem of the evolution of a sense of fairness similar to the problem of generalisation in evolutionary robotics, i.e. the problem to evolve robots that perform well in new contexts different from the ones used for evaluation.

- in humans, weak individuals could produce as much benefits as strong individuals, giving good outside options to weak individuals (see the discussion in Chapter 3).

- in humans, there could be more opportunities to cooperate in *small* groups. These opportunities allow an unconstrained partner choice because they allow weak individuals to avoid bullies.

- Bshary et al. (2002) argues that there could be a cumulative effect of cognitive skills, because contrarily to primates, no single fish species shows all kinds of social behaviors (i.e., all social behaviors found in primates can be found in fish but in different species, no fish species has them all):

  > *"We hypothesise that the diversity of skills found in individual primates as opposed to individual fish might reflect the major difference between the two taxa. Assuming that every additional skill needs an increase in neocortex size, the additive effects of computational power might even have led to fulguration (Lorenz 1973), the occurrence of a new system of traits that is not predictable from the traits themselves."*

  The emphasis here is on fish and primates but maybe similar "cumulative" effects can be found in humans compared to non-humans.

- van Schaik and Kappeler (2006) suggest that human cooperation differs from non-human primate cooperation on six points: in humans, there is more cooperation in groups as opposed to dyads, more cooperation with non-kins, more high-risk cooperation, existence of punishment, existence of reputation, and existence of trade.

- humans could have very long-term interactions, a necessary condition for a sense of fairness to be really useful. Long-lasting interactions exist also outside humans (for instance in fish, Croft et al. (2006)), so how long-term interactions need to be exactly is unclear.

Chapter 9

Three common misunderstandings

"I chose to keep most for myself, there were no penalties for not sharing."

AAS6RAD34TR

I have already talked a lot in the discussion sections of the previous chapters so I will try to be brief here. I just outline three common misunderstandings that I think important to clarify and, in the next chapter, I outline the four most interesting directions for research in my opinion. These two chapters are not an original research work, as many ideas are not mine and were already present in Baumard (2010); Sperber and Baumard (2012); Baumard et al. (2013); Baumard (2015), but they give more depth to this thesis and help to put it into perspective.

9.1 The role of reputation

One way to summarize our results is to say that "fairness evolves because fair individuals get a good reputation and get chosen as social partners". From there, it is easy to jump to the conclusion that fairness is about preserving one's reputation in a very self-interested manner: people would be fair only when they know that their reputation is at stake. This is not what we suggest, we suggest instead that concerns for fairness are automatic and genuine, and that concerns for fairness and concerns for reputation are separate at the proximate level. Maybe the best way to be convinced that concerns for fairness are distinct from concerns for reputation is to identify situations in which people go against the interests of their reputation for the sake of fairness. Examples might include a criminal who spontaneously walks into a police station, someone who confesses having cheated on her partner, or someone who prefers to lose a friend rather than turn a blind eye on something bad he has done. Conversely, someone not talking about something fair she has done, like giving a fair amount in a pay-what-you-want pricing system, could also be evidence

for this disconnection between care for fairness and care for reputation. (Giving a fair amount in anonymous economic games can also be interpreted as such evidence, but some authors will argue that subjects never really believe the interaction is anonymous).

That is not to say that we do not also manage our reputation, but this management would come from a different selection pressure than being chosen for *cooperative* activities. People like to defend their reputation in many other domains than the cooperative one. People like to have the reputation of being intelligent, being strong, being faithful, etc. People will often act to preserve those reputations but *without representing their actions as fair*. Those types of reputation can be useful in many situations that are not cooperative (for instance obtaining mates), and so there should definitely be a selection pressure to manage one's reputation. But if being strong or intelligent can be helpful in cooperative situations, it is not worth much if one is not also fair.

We are not saying neither that concerns for fairness necessarily prime over concerns for reputations. As the experimental literature has abundantly shown (Haley and Fessler, 2005), people will behave more or less prosocially depending on the presence of others or even just cues about the presence of others. At the psychological level, it is unclear to me whether this variability is directly produced by the sense of fairness (which would mean that the sense of fairness takes as input some reputational information, and is thus not entirely genuine) or whether this variability is the result of an inhibition of the sense of fairness by a reputation-management "module".

But if one agrees that concerns for fairness and concerns for reputation are two different things at the proximate level, our models do not help to explain this separation. The only thing that matters in our models is that individuals preserve their reputation, so a concern for reputation should in theory be enough to solve the evolutionary problem of being chosen as a social partner. More sophisticated versions of our models should be built to help clarify this point, but I can suggest a verbal explanation here. Concerns for fairness are distinct from concerns for reputation because of a problem of risk management: a reputation-management module will always make mistakes (make someone behave selfishly when she was actually observed), and those mistakes are very costly in the presence of reputation. Because of reputation, again not understood in the narrow sense of the opinion that I have of someone else, but in the broader sense of the socially transmitted judgment about someone that is widely shared (Sperber and Baumard, 2012), making a single mistake can cost one all of her cooperative partners. Mistakes are also easily de-

tected, because humans are expert mind-readers. When asking a favour to someone for instance, an hesitation of just a few seconds before answering is enough to be informed of someone's real motivations. Hence, because mistakes are easily detected and extremely costly, the best way to keep cooperative partners is to avoid making mistakes as much as possible and be fair consistently. As Sperber and Baumard (2012) put it, the best way to *appear* fair is to *be* fair. It would probably be worth it to formalize this idea in a model though, especially because the existence of genuine concerns for fairness is not easily accepted by many scholars (see section 1.5.5).

9.2 The role of emotions

Some researchers think fairness finds its roots in emotions like empathy, shame, anger, disgust, etc. These explanations present two problems. First, they are only proximal explanations, with no regard for the evolutionary reason of why those emotions did or did not evolved in humans and other species. Second, it is actually very difficult to find an emotion that always leads to fair behaviors. For instance, if someone gets hurt after being attacked it is true that we will feel empathy and find the situation unfair. But if someone cuts his finger while cooking, we will feel empathy without finding the situation unfair. The same is true for shame: if it is true that we can think of situations in which shame and fairness concerns coexist (after cheating on one's partner for instance), we also know situations in which they do not (feeling ashamed for forgetting to remove one's pyjamas before going out). The same is true for anger: if people witnessing unfairness often report being angry, anger often leads to overreactions that do not show the logic of proportionality between tort and punishment that should result from fairness. Hence, to postulate that fairness comes from emotions means that one needs a theory of why emotions lead to fair behaviors only in some cases and not others. This can hardly be done without postulating that a fairness judgement intervenes somewhere, and upstream from emotions in particular.

Hence, because emotions often lead us to behave in unfair ways, it is unlikely that fairness finds its source in "emotions" generally-speaking. Only guilt seems to always be associated with fair behaviors, and guilt has indeed been suggested to be an exclusively fairness-related emotion (Baumard et al., 2013). Guilt motivates us to repair our misdeeds, rather than hide them as would be the case with shame. But again it would not mean that guilt is at the origin of fairness, as we can identify unfair situations without necessarily feeling guilty (when others commit unfair actions). It only suggests that guilt is the best psychological trick that natural selection has found to make people behave in adaptive ways when they have done something

unfair.

9.3 The role of punishment

Is choosing to interact with fair individuals a punishment for unfair individuals? Probably not, at both the evolutionary and psychological level. At the evolutionary level, changing partners is often not costly for the actor. The abandoned partner might suffer a cost, but the behavior of changing partners has not evolved because it inflicted this cost, it does not have this function. Partner choice has evolved because it provided direct benefits to the cheated partners. A famous analogy in biology is the behavior of yucca-trees who allow yucca-moths to lay their eggs in their ovaries in exchange for pollination (Powell, 1992). But because the moth's larvae feed on the seeds, moths can cheat by putting too many eggs in the same ovary. When cheating happens, it has been shown that yucca-plants abort those ovaries, something that humans could interpret as a punishment. But, as Noë et al. (2001) puts it, "this is an economic decision, rather than a form of punishment".

Similarly, at the psychological level, it is unclear that people are trying to impose costs on their unfair partners so that they will behave fairly in the future. They are rather trying to cut their losses, with usually no expectation to interact with their unfair partners again in the future. If we agree that partner choice is not about punishing, it does not necessarily mean that costly punishment is not important in humans. But the present thesis has shown that partner choice alone was already sufficient to explain many aspects of fairness, so the additional explanatory power of punishment is unclear. In some fish species in which both partner choice and partner control are possible, fish seem to use partner control only when they do not have the possibility to choose partners (Bshary, 2001; Bshary and Noë, 2003), which could suggest that partner choice is a better strategy than partner control when the two are available.

It is possible that the importance of costly punishment for the evolution of human cooperation in the current literature is overestimated. Punishment is part of our daily life because of institutions, because of examples of punishment that are not linked to cooperation (retaliation for an adulterer for instance), and because in economic games people punish (maybe because they have nothing better to do, see also Chapter 7 on this point). But in small-scale societies, examples of costly punishment linked to cooperative behaviors are hard to find. In a review of the anthropological literature, Baumard (2010) showed that there is little evidence of punishment that can not be explained by retaliation (self-interested defense of one's

reputation, as many animals do) and almost no evidence for third-party punishment at all. Anthropologists in different traditional societies report that "the only painful result of anti-social actions was the loss of the esteem of others" (Radcliffe-Brown, 1922), "most disputes are resolved by self-segregation and attract hardly any attention" (Woodburn, 1982), "it is usual for one of the parties to leave the group, and the rivals will then avoid each others" (Furer-Haimendorf, 1967) (cited in Baumard, 2010). Johnson and Earle (2000) in their own review of human family-level societies agree that "a violation [of the common understandings concerning the proper support for one's kin and friends] is not a crime but an embarrassment; the violator is less likely to be physically punished than teased and ridiculed". Radcliffe-Brown (1922) also reports that "if one person injured another it was left to the injured one to seek vengeance if he wished and if he dared.", so if hunter-gatherers already hesitate to seek vengeance for cases as serious as physical injuries, it does not seem plausible to think that they happily punish those who have not shared or cooperated in the expected way.

Chapter 10

Interesting directions for research

"I'm not greedy. If it were the other way around, I'd hope my partner would split 50/50 with me."

AS2L45CS32LE

10.1 Generalized reciprocity

An important direction is, I think, to model the evolution of fairness in populations that match more closely what we know of hunter-gatherer populations. Now that we understand the basic partner choice mechanism that leads to the evolution of fairness, we should get rid of the populations of 1,000 individuals in which individuals are randomly drawn two-by-two to cooperate. Even if there is debate in anthropology as to the precise structure of modern and ancient hunter-gatherer societies, we could at least try to model populations structured in camps of 20-30 individuals whose composition can change from day to day, with family units moving between camps located in the same foraging area (Johnson and Earle, 2000; Dunbar, 1993). We should also try to model more diverse and realistic resources, for instance resources that are not easily divided (like safety, or child education) or resources that provide diminishing returns (Kaplan and Gurven, 2005; Nettle et al., 2011).

What I think these improvements will allow us to understand is the existence of what anthropologists call "generalized reciprocity" (Sahlins, 1972), a system in which food is shared without bookkeeping and precise expectations of reciprocity (see also section 4.5). Our models predict that the best producers should get more of the resource, but the data often contradict this prediction, showing more egalitarian sharings (Kaplan and Gurven, 2005; Gurven, 2004). There are lots of questions we could ask. Do egalitarian divisions result from the diminishing returns to the consumption of meat? Do good hunters keep smaller *fractions* of their food but

still get larger absolute quantities (Alvard, 2004)? Do they get delayed benefits by being chosen more in later cooperative interactions ("generalized equity")? Are good producers rewarded exactly in proportion to their contribution or are they happy with just receiving more than others? Is strict proportionality limited to interactions with people that are not part of the close social circle (Kaplan and Gurven, 2005)? Is generalized reciprocity about gaining prestige that has nothing to do with fairness (i.e. prestigious people could gain privileged access to women for instance but without the others thinking that they *deserve* those women)?

If there is a chance to answer some of these questions theoretically, it will only be through the introduction of more realism into the models. I find these questions all the more interesting that even us Westerners behave as "generalized reciprocators" sometimes, for instance when we go on a camping trip with strangers. We then agree to share benefits such as cooked food in an egalitarian way, even if some people made greater contributions to its production than others. Arguments can happen if someone always lies in the hammock while everybody is helping to prepare food, which shows that even in this egalitarian setting, we still monitor the fairness of our interactions. But arguments are more about how much people invest into cooperation (how much they help to prepare food), and not about how to divide the food. Nobody would think about asking for more baked potatoes as a way to compensate for their higher investment into cooking them. This hypothetical but I think plausible example for who is familiar with camping trips shows that even us Westerners, strongly influenced by meritocratic ideas, "automatically" turn to generalized reciprocity in some settings. This raises the interesting question of how many different ways to make an interaction fair there are, and what settings will favour one way over the other.

10.2 Fairness versus other motivations

If the sense of fairness is definitely an important aspect of human psychology, it is not the only one. Being chosen as a cooperative partner is only one of the many goals an individual has, with finding mates, have kids, have friends, find food, manage one's reputation (see section 9.1), etc. Those objectives can be conflicting, such as when someone defends her child even when she knows her child has done something unfair. Conflicting objectives are probably an important source of variability in human fair behaviors, other than differences in assessing contributions (discussed in section 4.5). A good illustration of this tension between fairness and selfishness is the justification of one of the subjects in my video game experiment: *"The other guy did a good job and I felt like it was fair, but I couldn't help taking just a little*

bit more than him (10%)". But what situations will make fairness prevail and what situations will make selfishness prevail is largely unknown.

Another interesting competing motivation is competitive altruism. One of the important results of this thesis is to show that partner choice in a biological market does not necessarily lead to competitive altruism (in the exaggerated sense of no-limit runaway generosity, see Chapter 6). Competitive altruism and fairness could coexist in the wild (because they are responses to different selection pressures) but their relative importance is unclear. It seems to me that in natural settings, people prefer fair people over generous people, as over-generosity is often seen suspiciously (Bird and Power, 2015). In a similar vein, Sperber and Baumard (2012) discuss why Mother Theresa types are rare in humans:

> *"However much you may admire a saintly person, she might not be your first choice as a partner in cooperation: her giving too much and asking too little would put your more balanced behaviour in a bad light and might cause you to feel embarrassed. Her duties being to humankind (or to god), she may at any time leave you flat in order to achieve a greater good."*

Striking anthropological examples of refusals of large gifts also show that considerations of fairness, i.e. acknowledgements that gifts create duties, prime on considerations of generosity in certain situations.

On the other hand, food sharing in generalized reciprocity (that is not correlated with contributions) could be interpreted as an instance of competitive altruism (but see section 4.5, 10.1 and Baumard et al., 2013 for how they can also be interpreted as intances of fairness). It is thus unclear how competitive altruism and fairness work out together in natural settings. In Barclay (2013)'s review on "strategies for cooperation in biological markets, especially in humans", I could not help but notice that fairness is not mentioned as one such strategy, and not mentioned at all in the paper. Maybe this is just a terminological issue (fairness might be seen as a special case of competitive altruism), but I hope that the present thesis will have shown that because fairness is not about being generous but about balancing the costs and benefits of being generous, it deserves a special place as a particularly elegant solution found by natural selection to behave in a biological market.

10.3 Market situations that are deemed unfair

There are some market situations that people find unfair. The most famous example comes from Kahneman et al. (1986b):

A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store raises the price to $20. Please rate this action as: Completely Fair / Acceptable / Unfair / Very Unfair

82% of respondents find it either unfair or very unfair. This result can be seen as a problem for our theory. If the store can raise its price, it is because it has just gained a better bargaining position in the market, in a similar way that talented people have better outside options in a biological market. So, if as we suggest, fairness is the result of market mechanisms at the ultimate level, why are humans not finding this market situation fair? Why do we find fair to give more to the best producers but not to the hardware store, if all that matters are market mechanisms?

We can suggest some hypotheses. Increasing prices after a snowstorm reveals that one is more interested in short-term profits rather than long-term benefits. People interested in short-term profits are not the kind of people we should be looking for to cooperate with, as they will be ready to abandon us at the first occasion. There is also a difference between a talented person and the hardware store because the talented person *always* has good outside options throughout his life, whereas the store saw an increase of outside options that was (i) only punctual and (ii) not earned through effort.

Ideally, those hypotheses should be formalized in a model. It is not too hard to make up explanations for why judging free-market outcomes as unfair is adaptive, but how those proximate judgements came to be from ultimate market mechanisms is unclear. Investigating this issue is a project that I considered at the beginning of my thesis but that I abandoned later in favour of other projects, so I am looking forward to seeing more research on this subject.

10.4 How much of morality can fairness explain?

For fairness to be an effective partner choice criterion, it has to assess costs and benefits in many different ways. Said differently, fairness is only as good as its measurements of costs and benefits are. Seeing the diversity of situations in which humans cooperate, it is probable that this requires costs and benefits to be measured in a very abstract way, and that they should take into account way more than just the time invested and the amount of resources produced. To be more concrete, we can go back to the examples given in the introduction. When people ask whether it is fair to keep their cat indoors, there are no considerations of effort or productivity here. But we can still fruitfully interpret this situation in terms of costs and benefits: people are asking whether the costs for the cat not to have access to outdoors are

too high to be compensated by the benefits to the cat obtained through the good care of the cat's owner. The same is true when people ask whether it is fair to have a baby at 40, or to use legendary pokemons: it is not always easy to know what costs are computed exactly but they must be in a way or another, and can not be based on computations of effort or productivity only.

The cat example already shows that fairness judgements can happen outside distributive and cooperative situations. But if we accept that costs and benefits are automatically computed in any situation, maybe fairness judgements also intervene in situations that people categorize as moral rather than fair. A legitimate question is thus: how much of morality can fairness explain?

I suggest to take a concrete example, with the most famous experiment in moral psychology, the trolley dilemma (Thomson, 1985). In a first version of this experiment, subjects are confronted to a hypothetical scenario in which a trolley is rushing at 5 people lying on the track. The subject has to decide whether or not to deviate the trolley on a secondary track, which will save the 5 people but will kill one person lying on this secondary track. In this scenario, a robust result is that a majority of people accept to deviate the trolley. Nonetheless, in a second version of this experiment, one needs to push a bystander on the track to stop the trolley (there is no secondary track). In this case, a majority of people refuse to do so. This can seem surprising as the benefits of killing one person are the same in the two versions. Scholars have advanced several explanations to account for those results, for instance that morality is based on the principle of "double effect" (it is wrong to use someone as a mean), or that applying a *direct* force to inflict harm is always wrong (Greene, 2013). Independently of how true those explanations are, it is important to see that there are also costs and benefits distributed in the trolley dilemma (Baumard, 2015): subjects are asked to allocate death (or life) to different people. And it is possible to see why allocating death to the single person is less costly in the first than in the second version of the dilemma. When someone is already lying on the tracks, she is not safe at all: in a colourful sense, she is already "close" from death. On the contrary, when someone has to be pushed to be put on the tracks, it means she was in a safer position, "further from death". The costs inflicted to the single person in the process of killing her are thus higher in the second than in the first version (Baumard, 2015), which could explain why people find killing in the second version more immoral - or, maybe, more unfair.

The hypothesis that moral judgments have a lot to do with fairness makes predictions that can be tested. In the case of the trolley dilemma, it predicts that manipulating the degree of safety of the single person (and thus the costs associ-

ated with killing her) will affect people's moral judgement about the situation. I am looking forward to experiments testing this prediction. At this point it is too early to say how much of morality (understood in a broad sense and encompassing situations like respect of hierarchical relationships, issues of purity and impurity...) fairness can explain. But it is to me the most interesting empirical investigation to undertake, and the one that will increase the most our understanding of morality.

Part VI

Conclusion

Some of the results I have presented here would probably be trivial for an economist working on markets. Obtaining proportional outcomes as a result of each individual making sure to receive the same return on investment in each interaction (Chapter 4) might be one such result. The fact that physically weaker individuals do not necessarily have bad outside options in a biological market (Chapter 3) might have been less remarked. In any case, our main contribution is to draw attention to how useful it is to locate those market mechanisms at the ultimate level in order to explain the behavioral data at hand. The fact that so many competing explanations persist in the literature (Chapter 1) clearly shows that many people do not find it obvious to link fairness with market mechanisms, whether ultimate or proximate. We should not forget neither that explanations of prosocial behaviors based on cultural group selection are predominant in many fields today, whereas our explanation relies on individual selection only. If we are right, the sense of fairness is "simply" the product of market mechanisms operating at the ultimate level. If we are right, the sense of fairness is natural selection's way to help humans navigate their social world in which they have been constantly choosing and being chosen as cooperative partners. As I was not the first one to suggest this hypothesis, I hope to remain modest by saying that I find this hypothesis remarkably simple and powerful at the same time. Even in my day-to-day life, I am surprised at how much of people's outraged behaviors I can understand by adopting a costs and benefits perspective. I will be happy if I have contributed to push this theory forward if only a little, and I am excited to see where it will be twenty years from now, in particular in explaining the Holy Grail of many philosophers - human morality.

Bibliography

Adair, J. G., 1984. The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology 69:334–345.

Adams, J. S., 1963. Toward an Understanding of Inequity. Journal of abnormal psychology 67:422–436.

Adams, J. S. and P. R. Jacobsen, 1964. Effects of Wage Inequities on Work Quality. Journal of abnormal psychology 69:19–25.

Aktipis, C. A., 2004. Know when to walk away: contingent movement and the evolution of cooperation. Journal of theoretical biology 231:249–60.

Aktipis, C. A., 2011. Is cooperation viable in mobile organisms? Simple Walk Away rule favors the evolution of cooperation in groups. Evolution and Human Behavior 32:263–276.

Alexander, J., 2007. The Structural Evolution of Morality. Cambridge University Press.

Alvard, M., 2004. Good hunters keep smaller shares of larger pies. Comment on To give and to give not: the behavioral ecology of human food transfers by Michael Gurven. Behavioral and Brain Sciences 27:560–561.

Alvard, M. S., 2002. Carcass ownership and meat distribution by big-game cooperative hunters, vol. 21.

Alvard, M. S. and D. a. Nolin, 2002. Rousseau ' s Whale Hunt ? Current Anthropology 43:533–559.

Amici, F., E. Visalberghi, and J. Call, 2014. Lack of prosociality in great apes , capuchin monkeys and spider monkeys : convergent evidence from two different food distribution tasks. Proc. R. Soc.B .

Amir, O., D. G. Rand, and Y. K. Gal, 2012. Economic games on the internet: The effect of $1 stakes. PLoS ONE 7:1–4.

André, J.-B., 2014. Mechanistic constraints and the unlikely evolution of reciprocal cooperation. Journal of evolutionary biology Pp. 1–12.

André, J.-B. and N. Baumard, 2011a. Social opportunities and the evolution of fairness. Journal of theoretical biology 289:128–35.

André, J.-B. and N. Baumard, 2011b. The evolution of fairness in a biological market. Evolution 65:1447–56.

Arak, A. and M. Enquist, 1993. Hidden preferences and the evolution of signals. Philosophical transactions of the Royal Society B. 340:207–213.

Aristotle, 1999. Nicomachean Ethics, vol. 112.

Arnott, G. and R. W. Elwood, 2009. Assessment of fighting ability in animal contests. Animal Behaviour 77:991–1004.

Austin, W. and E. Walster, 1974. Reactions to confirmations and disconfirmations of expectancies of equity and inequity.

Bailey, R., 1991. The behavioral ecology of Efe Pygmy men in the Ituri Forest, Zaire. University of Michigan Museum.

Barclay, P., 2004. Trustworthiness and competitive altruism can also solve the "tragedy of the commons". Evolution and Human Behavior 25:209–220.

Barclay, P., 2006. Reputational benefits for altruistic punishment. Evolution and Human Behavior 27:325–344.

Barclay, P., 2011. Competitive helping increases with the size of biological markets and invades defection. Journal of theoretical biology 281:47–55.

Barclay, P., 2013. Strategies for cooperation in biological markets, especially for humans. Evolution and Human Behavior .

Barclay, P. and B. Stoller, 2014. Local competition sparks concerns for fairness in the ultimatum game. Biology letters .

Barclay, P. and R. Willer, 2007. Partner choice creates competitive altruism in humans. Proceedings. Biological sciences / The Royal Society 274:749–53.

Bardsley, N., 2008. Dictator game giving: Altruism or artefact? Experimental Economics 11:122–133.

Barker, J. L., P. Barclay, and H. K. Reeve, 2012. Within-group competition reduces cooperation and payoffs in human groups. Behavioral Ecology 23:735–741.

Barkow, J. H., L. Cosmides, and J. Tooby, 1992. The adapted mind: evolutionary psychology and the generation of culture. Oxford University Press.

Bateson, M., D. Nettle, and G. Roberts, 2006. Cues of being watched enhance cooperation in a real-world setting. Biology letters 2:412–4.

Baumard, N., 2010. Has punishment played a role in the evolution of cooperation? A critical review. Mind & Society 9:171–192.

Baumard, N., 2015. The Origins of Fairness: How Evolution Explains Our Moral Nature. Oxford University Press.

Baumard, N., J. André, and D. Sperber, 2013. A mutualistic approach to morality: The evolution of fairness by partner choice. Behavioral and Brain Sciences 6:59–122.

Baumard, N. and P. Boyer, 2013. Explaining moral religions. Trends in cognitive sciences 17:272–80.

Baumard, N., O. Mascaro, and C. Chevallier, 2012. Preschoolers are able to take merit into account when distributing goods. Developmental psychology 48:492–498.

Baumard, N. and D. Sperber, 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. and P. Tompkinson, 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. and L. Samuelson, 1994. An economist's perspective on the evolution of norms. Journal of Institutional and Theoretical Economics . . . 150:45–63.

Bird, R. B. and E. a. Power, 2015. Prosocial signaling and cooperation among Martu hunters. Evolution and Human Behavior .

Boehm, C., 1993. Egalitarian behavior and reverse dominance hierarchy. Current Anthropology 34:227–254.

Boehm, C., 1997. Impact of the human egalitarian syndrome on Darwinian selection mechanics. American Naturalist 150:S100–S121.

Bongard, J. C., 2013. Evolutionary Robotics. Communications of the ACM 56.

Boyd, R., 2006. Reciprocity: you have to think different. Journal of evolutionary biology 19:1380–1382.

Bräuer, J., J. Call, and M. Tomasello, 2006. Are apes really inequity averse? ... of the Royal ... Pp. 3123–3128.

Bräuer, J. and D. Hanus, 2012. Fairness in Non-human Primates? Social Justice Research 25:256–276.

Brosnan, S., H. C. Schiff, and F. B. M. de Waal, 2005. Tolerance for inequity may increase with social closeness in chimpanzees. Proc. R. Soc.B Pp. 253–258.

Brosnan, S. and F. de Waal, 2014. Evolution of responses to (un) fairness. Science .

Brosnan, S. F., 2006. Nonhuman species' reactions to inequity and their implications for fairness, vol. 19.

Brosnan, S. F., 2013. Justice- and fairness-related behaviors in nonhuman primates. Proceedings of the National Academy of Sciences 2013.

Brosnan, S. F. and F. B. M. De Waal, 2003. Monkeys reject unequal pay. Nature 425:297–299.

Brosnan, S. F., C. Talbot, M. Ahlgren, S. P. Lambeth, and S. J. Schapiro, 2010. Mechanisms underlying responses to inequitable outcomes in chimpanzees, Pan troglodytes. Animal Behaviour 79:1229–1237.

Brosnan, S. F. and F. B. M. de Waal, 2012. Fairness in Animals: Where to from Here? Social Justice Research 25:336–351.

Bshary, R., 2001. The cleaner fish market. *in* Economics in nature: social dilemmas, mate choice and biological markets.

Bshary, R. and R. Noë, 2003. Biological Markets: The ubiquitous influence of partner choice on the dynamics of cleaner fish Client reef fish interactions. Genetic and Cultural Evolution of Cooperation 9:167–184.

Bshary, R. and N. J. Raihani, 2013. "Fair" outcomes without morality in cleaner wrasse mutualism (response to BBS paper "A mutualistic approach to morality: The evolution of fairness by partner choice"). Behavioral and brain sciences 6.

Bshary, R. and D. Schäffer, 2002. Choosy reef fish select cleaner fish that provide high-quality service. Animal Behaviour 63:557–564.

Bshary, R., W. Wickler, and H. Fricke, 2002. Fish cognition: a primate's eye view. Animal cognition 5:1–13.

Burkart, J. M., O. Allon, F. Amici, C. Fichtel, C. Finkenwirth, a. Heschl, J. Huber, K. Isler, Z. K. Kosonen, E. Martins, E. J. Meulman, R. Richiger, K. Rueth, B. Spillmann, S. Wiesendanger, and C. P. van Schaik, 2014. The evolutionary origin of human hyper-cooperation. Nature communications 5:4747.

Burkart, J. M., E. Fehr, C. Efferson, and C. P. van Schaik, 2007. Other-regarding preferences in a non-human primate: common marmosets provision food altruistically. Proceedings of the National Academy of Sciences of the United States of America 104:19762–19766.

Call, J. and M. Tomasello, 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 12:187–192.

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

Cappelen, A., A. Hole, E. Sø rensen, and B. Tungodden, 2007. The pluralism of fairness ideals: An experimental approach. American Economic Review 97:818–827.

Cappelen, A., E. Sø rensen, and B. Tungodden, 2010. Responsibility for what? Fairness and individual responsibility. European Economic Review 54:429–441.

Cappelen, A. W., T. Halvorsen, E. O. Sø rensen, and B. Tungodden, 2013. Face-saving or fair-minded: What motivates moral behavior? .

Carlsson, F., H. He, and P. Martinsson, 2013. Easy come, easy go: The role of windfall money in lab and field experiments. Experimental Economics 16:190–207.

Cashdan, E., 1980. Egalitarianism among hunters and gatherers. American Anthropologist 82:116–120.

Cason, T. and A. Williams, 1990. Competitive equilibrium convergence in a posted-offer market with extreme earnings inequities. Journal of Economic Behavior & Organization 14:331–352.

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

Charnov, E. L., 1976. Optimal foraging, the marginal value theorem. Theoretical population biology 9:129–136.

Chase, V. M., R. Hertwig, and G. Gigerenzer, 1998. Visions of rationality. Trends in Cognitive Sciences 2:206–214.

Chen, M. K. and L. R. Santos, 2006. Some thoughts on the adaptive function of inequity aversion: An alternative to Brosnan's social hypothesis. Social Justice Research 19:201–207.

Cherry, T. L., P. Frykblom, and J. F. Shogren, 2002. Hardnose the dictator. American Economic Review 92:1218–1221.

Chevallier, C., J. Xu, K. Adachi, J.-B. van der Henst, and N. Baumard, 2015. Preschoolers' Understanding of Merit in Two Asian Societies. Plos One 10:e0114717.

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.

Chiang, Y.-S., 2010. Self-interested partner selection can lead to the emergence of fairness. Evolution and Human Behavior 31:265–270.

Christen, M. and H. J. Glock, 2012. The (Limited) Space for Justice in Social Animals. Social Justice Research 25:298–326.

Cohen, G., 2009. Why not socialism? Princeton University Press.

Croft, D. P., R. James, P. O. R. Thomas, C. Hathaway, D. Mawdsley, K. N. Laland, and J. Krause, 2006. Social structure and co-operative interactions in a wild population of guppies (Poecilia reticulata). Behavioral Ecology and Sociobiology 59:644–650.

Cuzick, J., 1985. A Wilcoxon-Type Test for Trend. Statistics in Medicine 4:543–547.

Dawes, C. T., J. H. Fowler, T. Johnson, R. Mcelreath, and O. Smirnov, 2007. Egalitarian motives in humans. Nature 446:794–796.

Debove, S., J.-b. Andre, and N. Baumard, 2015a. Partner choice creates fairness in humans. Proc. R. Soc.B 282.

Debove, S., N. Baumard, and J.-B. André, 2015b. Evolution of equal division among unequal partners. Evolution 69:561–569.

Dindo, M. and F. B. M. De Waal, 2007. Partner effects on food consumption in brown capuchin monkeys. American Journal of Primatology 69:448–456.

Doncieux, S., J.-b. Mouret, N. Bredeche, and V. Padois, 2011. Evolutionary Robotics : Exploring New Horizons. Pp. 3–25, *in* New Horizons in Evolutionary Robotics.

Dreber, A., T. Ellingsen, M. Johannesson, and D. G. Rand, 2013. Do people care about social context? Framing effects in dictator games. Experimental Economics 16:349–371.

Duan, W.-Q. and H. E. Stanley, 2010. Fairness emergence from zero-intelligence agents. Physical Review E 81:026104.

Dugatkin, L. A., 1997. Cooperation among animals: an evolutionary perspective. Oxford University Press.

Dugatkin, L. A. and D. S. Wilson, 1993. Fish behaviour, partner choice experiments and cognitive ethology. Reviews in Fish Biology and Fisheries 3:368–372.

Dunbar, R. I., 1993. Coevolution of neocortical size, group size and language in humans. Behavioral and Brain Sciences 16:681–735.

Eckel, C. C. and P. J. Grossman, 1996. Altruism in Anonymous Dictator Games. Games and Economic Behavior 16:181–191.

Ellis, L., 1995. Dominance and Reproductive Success Among Nonhuman Animals: A Cross-Species Comparison. Ethology and sociobiology 333:257–333.

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

Enquist, M. and A. Arak, 1994. Symmetry, beauty and evolution. Nature .

Ezoe, H. and Y. Iwasa, 1997. Evolution of condition-dependent dispersal: A genetic-algorithm search for the ESS reaction norm. Researches on population ecology 39:127–137.

Fairtrade International, 2014. Monitoring the Scope and Benefits of Fairtrade - sixth edition 2014. Tech. rep.

Fehr, E., H. Bernhard, and B. Rockenbach, 2008. Egalitarianism in young children. Nature 454:1079–1084.

Fehr, E. and U. Fischbacher, 2003. The nature of human altruism. Nature 425:785–91.

Fehr, E. and S. Gächter, 2002. Altruistic punishment in humans. Nature 415:137–40.

Fehr, E. and K. Schmidt, 1999. A theory of fairness, competition, and cooperation. The quarterly journal of economics 114:817–868.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10:171–178.

Fischbacher, U., C. M. Fong, and E. Fehr, 2009. Fairness, errors and the power of competition. Journal of Economic Behavior & Organization 72:527–545.

Fiske, A. P., 1992. The four elementary forms of sociality: framework for a unified theory of social relations. Psychological review 99:689–723.

Floreano, D. and C. Mattiussi, 2008. Bio-Inspired Artificial Intelligence: theories, methods, and technologies.

Forber, P. and R. Smead, 2014. The evolution of fairness through spite. Proceedings of the Royal Society B 281.

Forsythe, R. and J. Horowitz, 1994. Fairness in simple bargaining experiments. Games and economic behavior 6.

Franzen, A. and S. Pointner, 2012. The external validity of giving in the dictator game. Experimental Economics Pp. 155–169.

Frohlich, N., J. Oppenheimer, and J. Bernard Moore, 2001. Some doubts about measuring self-interest using dictator experiments: The costs of anonymity. Journal of Economic Behavior and Organization 46:271–290.

Frohlich, N., J. Oppenheimer, and A. Kurki, 2004. Modeling other-regarding preferences and an experimental test. Public Choice 119:91–117.

Fruteau, C., B. Voelkl, E. van Damme, and R. Noë, 2009. Supply and demand determine the market value of food providers in wild vervet monkeys. Proceedings of the National Academy of Sciences of the United States of America 106:12007–12012.

Fudenberg, D. and E. Maskin, 1986. The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54.

Furer-Haimendorf, C., 1967. Morals and merit: a study of values and social controls in South Asian societies. Weidenfeld & Nicolson, London Gintis.

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

Gardner, a. and S. a. West, 2004. Spite and the scale of competition. Journal of evolutionary biology 17:1195–203.

Brañas Garza, P., 2007. Promoting helping behavior with framing in dictator games. Journal of Economic Psychology 28:477–486.

Gavrilets, S., 2012. On the evolutionary origins of the egalitarian syndrome. Proceedings of the National Academy of Sciences of the United States of America 109:14069–74.

Gavrilets, S. and S. M. Scheiner, 1993. The genetics of phenotypic of reaction norm shape V . Evolution of reaction norm shape. Journal of evolutionary biology 48:31–48.

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

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

Grafen, A., 1987. The logic of divisively asymmetric contests: respect for ownership and the desperado effect. Animal Behaviour 35:462–467.

Greene, J., 2013. Moral tribes. London: Penguin Press.

Grosskopf, B., 2003. Reinforcement and directional learning in the ultimatum game with responder competition. Experimental Economics 158:141–158.

Gurven, M., 2004. To give and to give not: The behavioral ecology of human food transfers. Behavioral and Brain Sciences .

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

Güth, W., N. Marchand, and J. Rullière, 1998. Equilibration et dépendance du contexte. Une évaluation expérimentale du jeu de négociation sous ultimatum [Equilibration and context dependency : an experimental investigation of the ultimatum bargaining game]. Revue économique 49:785–794.

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

Hagel, J. H. and A. E. Roth, 1995. Handbook of experimental economics. Princeton University Press, Princeton, New Jersey.

Haley, K. J. and D. M. Fessler, 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 Pp. 193–205.

Hammerstein, P. and G. a. Parker, 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 Pp. 611–633.

Harms, W., 1997. Evolution and ultimatum bargaining. Theory and decision Pp. 147–175.

Hart, B. L. and L. a. Hart, 1992. Reciprocal allogrooming in impala, Aepyceros melampus. Animal Behaviour 44:1073–1083.

Hatfield, G., 2005. Introspective Evidence in Psychology. Pp. 259–286, *in* Scientific Evidence: Philosophical Theories & Applications.

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

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, Others, P. M. Todd, V. M. Chase, R. Hertwig, G. Gigerenzer, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton, and D. Tracer, 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. The Behavioral and brain sciences 28:795–815; discussion 815–55.

Henrich, J., S. J. Heine, and A. Norenzayan, 2010. The weirdest people in the world? Behavioral and brain sciences 33:61–83.

Hill, K., 2002. Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. Human Nature 13:105–128.

Hoebel, E. A., 1954. The Law of Primitive Man: A Study in Comparative Legal Dynamics P. 372.

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

Hoffman, E., K. McCabe, K. Shachat, and V. Smith, 1994. Preferences, property rights, and anonymity in bargaining games. Games and Economic Behavior .

Homans, G. C., 1958. Social Behavior as Exchange. The american journal of sociology 63:597–606.

Hooper, P. L., M. Gurven, and H. Kaplan, 2014. Social and Economic Underpinnings of Human Biodemography. *in* M. Weinstein and M. A. Lane, eds. Sociality, Hierarchy, Health: Comparative Biodemography: Papers from a Workshop, Pp. 169–195. National Academies Press (US).

Horner, V., J. D. Carter, M. Suchak, and F. B. M. de Waal, 2011. Spontaneous prosocial choice by chimpanzees. Proceedings of the National Academy of Sciences of the United States of America 108:13847–13851.

Horowitz, A., 2012. Fair is Fine, but More is Better: Limits to Inequity Aversion in the Domestic Dog. Social Justice Research 25:195–212.

Howitt, D. and J. Mccabe, 1978. Attitudes do predict behaviour - in mails at least. British Journal of Social and Clinical Psychology 17:285–286.

Huck, S. and J. Oechssler, 1999. The Indirect Evolutionary Approach to Explaining Fair Allocations. Games and Economic Behavior 28:13–24.

Ichinose, G., 2012. Coevolution of Role Preference and Fairness. Complexity 00:1–9.

Ichinose, G. and H. Sayama, 2014. Evolution of fairness in the not quite ultimatum game. Scientific reports 4:5104.

Ipeirotis, P., 2010. Demographics of Mechanical Turk. NYU Working Paper No. CEDER-10-01 .

Iranzo, J., L. M. Floría, Y. Moreno, and A. Sánchez, 2012. Empathy emerges spontaneously in the ultimatum game: small groups and networks. PloS one 7:e43781.

Iranzo, J., J. Román, and A. Sánchez, 2011. The spatial Ultimatum game revisited. Journal of theoretical biology 278:1–10.

Jensen, K., J. Call, and M. Tomasello, 2007. Chimpanzees are rational maximizers in an ultimatum game. Science (New York, N.Y.) 318:107–109.

Jensen, K., B. Hare, J. Call, and M. Tomasello, 2006. What ' s in it for me ? Self-regard precludes altruism and spite in chimpanzees Pp. 1013–1021.

Johnson, A. W. and T. Earle, 2000. The Evolution of Human Societies: From Foraging Group to Agrarian State, Second Edition. Stanford University Press.

Johnstone, R. and R. Bshary, 2008. Mutualism, market effects and partner control. Journal of evolutionary biology 21:879–888.

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

Kahneman, D., J. Knetsch, and R. Thaler, 1986a. Fairness as a constraint on profit seeking: Entitlements in the market. The American economic review 76:728–741.

Kahneman, D., J. Knetsch, and R. Thaler, 1986b. Fairness as a constraint on profit seeking: Entitlements in the market. The American economic review 76:728–741.

Kahneman, D., J. L. Knetsch, and R. H. Thaler, 1986c. Fairness and the Assumptions of Economics. The Journal of Business 59:S285.

Kanngiesser, P., N. Gjersoe, and B. M. Hood, 2010. The effect of creative labor on property-ownership transfer by preschool children and adults. Psychological science 21:1236–41.

Kaplan, H. and M. Gurven, 2005. The Natural History of Human Food Sharing and Cooperation : A Review and a New Multi-Individual Approach to the Negotiation of Norms. Pp. 75–113, *in* R. B. . E. F. H. Gintis, S. Bowles, ed. Moral sentiments and material interests: The foundations of cooperation in economic life. MIT Press.

Kaplan, H. S., P. L. Hooper, and M. Gurven, 2009. The evolutionary and ecological roots of human social organization. Philosophical transactions of the Royal Society B. 364:3289–99.

Killingback, T. and E. Studer, 2001. Spatial Ultimatum Games, collaborations and the evolution of fairness. Proceedings. Biological sciences / The Royal Society 268:1797–801.

Kim, J.-Y., M. Natter, and M. Spann, 2009. Pay What You Want: A New Participative Pricing Mechanism. Journal of Marketing 73:44–58.

Kirchsteiger, G., 1994. The role of envy in ultimatum games. Journal of economic behavior & organization 2681.

Knez, M. and C. Camerer, 1995. Outside options and social comparison in three-player ultimatum game experiments. Games and Economic Behavior .

Knoch, D., A. Pascual-Leone, K. Meyer, V. Treyer, and E. Fehr, 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science (New York, N.Y.) 314:829–32.

Konow, J., 2000. Fair shares: Accountability and cognitive dissonance in allocation decisions. The American Economic Review 90:1072–92.

Konow, J., 2003. Which is the fairest one of all? A positive analysis of justice theories. Journal of economic literature XLI:1188–1239.

Kymlicka, W., 2002. Contemporary Political Philosophy - An Introduction. Oxford University Press, USA.

Leary, M. and R. Kowalski, 1990. Impression management: A literature review and two-component model. Psychological bulletin 107:34–47.

Lehmann, L., K. Bargum, and M. Reuter, 2006. An evolutionary analysis of the relationship between spite and altruism. Journal of evolutionary biology 19:1507–16.

Lehmann, L. and L. Keller, 2006. The evolution of cooperation and altruism - A general framework and a classification of models. Journal of Evolutionary Biology 19:1365–1376.

Levitt, S. D. and J. A. List, 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? The Journal of Economic Perspectives 21:153–174.

Liénard, P., C. Chevallier, O. Mascaro, P. Kiura, and N. Baumard, 2013. Early understanding of merit in Turkana children. Journal of Cognition and Culture 13:57–66.

List, J., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115:482–493.

Lyle, H. F. and E. a. Smith, 2014. The reputational and social network benefits of prosociality in an Andean community. Proceedings of the National Academy of Sciences of the United States of America 111:4820–5.

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

Massera, G., T. Ferrauto, O. Gigliotta, and S. Nolfi, 2014. Designing adaptive humanoid robots through the FARSA open-source framework. Adaptive Behavior 22:255–265.

Maynard Smith, J. and G. Parker, 1976. The logic of asymmetric contests. Animal Behaviour Pp. 159–175.

Maynard Smith, J. and G. Price, 1973. The logic of animal conflict. Nature 246.

McNamara, J., Z. Barta, L. Fromhage, and A. Houston, 2008. The coevolution of choosiness and cooperation. Nature 451:189–192.

Melis, A. A. P., B. Hare, and M. Tomasello, 2006. Engineering cooperation in chimpanzees: tolerance constraints on cooperation. Animal Behaviour 72:275–286.

Melis, A. P. and D. Semmann, 2010. How is human cooperation different? Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365:2663–2674.

Mellers, B. a., 1982. Equity judgment: A revision of Aristotelian views. Journal of Experimental Psychology: General 111:242–270.

Muthoo, A., 1999. Bargaining Theory with Applications. Cambridge University Press, Cambridge, England.

Nash, J., 1950. The bargaining problem. Econometrica 18:155–162.

Nesse, R. M., 2007. Runaway social selection for displays of partner value and altruism. Biological Theory 2.

Nettle, D., K. Panchanathan, T. S. Rai, and A. P. Fiske, 2011. The Evolution of Giving, Sharing, and Lotteries. Current Anthropology 52:747–756.

Noë, R. and P. Hammerstein, 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. and P. Hammerstein, 1995. Biological markets. Trends in ecology & evolution 10:336–9.

Noë, R., J. V. Hooff, and P. Hammerstein, 2001. Economics in nature: social dilemmas, mate choice and biological markets. Cambridge University Press, Cambridge, England.

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

Novakova, J. and J. Flegr, 2013. How Much Is Our Fairness Worth? The Effect of Raising Stakes on Offers by Proposers and Minimum Acceptable Offers in Dictator and Ultimatum Games. PLoS ONE 8.

Nowak, M. a., K. M. Page, and K. Sigmund, 2000. Fairness versus reason in the ultimatum game. Science 289:1773–5.

Orne, M., 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist 17:776–783.

Osborne, M. and A. Rubinstein, 1990. Bargaining and markets. Academic Press, Inc, San Diego, California.

Oxoby, R. J. and J. Spraggon, 2008. Mine and yours: Property rights in dictator games. Journal of Economic Behavior & Organization 65:703–713.

Page, K. M. and M. a. Nowak, 2001. A generalized adaptive dynamics framework can describe the evolutionary Ultimatum Game. Journal of theoretical biology 209:173–9.

Page, K. M. and M. a. Nowak, 2002. Empathy leads to fairness. Bulletin of mathematical biology 64:1101–16.

Page, K. M., M. a. Nowak, and K. Sigmund, 2000. The spatial ultimatum game. Proceedings. Biological sciences / The Royal Society 267:2177–82.

Piketty, T. and E. Saez, 2014. Inequality in the long run. Science 344.

Powell, J. A., 1992. Interrelationships of Yuccas and Yucca Moths. Trends in ecology & evolution 7.

Proctor, D., R. a. Williamson, F. B. M. D. Waal, and S. F. Brosnan, 2012. Chimpanzees play the ultimatum game. Proceedings of the National Academy of Sciences Pp. 1–6.

Radcliffe-Brown, A., 1922. The Andaman islanders: a study in social anthropology. The University press, Cambridge.

Raihani, N. J., R. Mace, and S. Lamba, 2013. The Effect of $1, $5 and $10 Stakes in an Online Dictator Game. PLoS ONE 8:3–8.

Raihani, N. J. and K. McAuliffe, 2012. Does Inequity Aversion Motivate Punishment? Cleaner Fish as a Model System. Social Justice Research 25:213–231.

Raihani, N. J., K. McAuliffe, S. F. Brosnan, and R. Bshary, 2012. Are cleaner fish, Labroides dimidiatus, inequity averse? Animal Behaviour 84:665–674.

Rand, D. G., J. D. Greene, and M. a. Nowak, 2012. Spontaneous giving and calculated greed. Nature 489:427–30.

Rand, D. G., C. E. Tarnita, H. Ohtsuki, and M. a. Nowak, 2013. Evolution of fairness in the one-shot anonymous Ultimatum Game. Proceedings of the National Academy of Sciences of the United States of America 110:2581–6.

Range, F., L. Horn, Z. Viranyi, and L. Huber, 2009. The absence of reward induces inequity aversion in dogs. Proceedings of the National Academy of Sciences of the United States of America 106:340–345.

Range, F., K. Leitner, and Z. Virányi, 2012. The Influence of the Relationship and Motivation on Inequity Aversion in Dogs. Social Justice Research 25:170–194.

Reeve, H., S. Emlen, and L. Keller, 1998. Reproductive sharing in animal societies: reproductive incentives or incomplete control by dominant breeders? Behavioral Ecology 9:267–278.

Ridley, M., 2004. Evolution. Blackwell Publishing, Malden, USA.

Roberts, G., 1998. Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society B: Biological Sciences 265:427–431.

Robinson, P. H. and R. Kurzban, 2007. Concordance and Conflict in Intuitions of Justice. Minn. L. Rev. 91:1–75.

Ross, J., A. Zaldivar, L. Irani, B. Tomlinson, and S. Silberman, 2010. Who are the Crowdworkers ? Shifting Demographics in Mechanical Turk. Proceeding CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems Pp. 2863–2872.

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

Roth, A. and V. Prasnikar, 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. The American Economic Review 81:1068–1095.

Rousseau, J.-J., 1762. Of The Social Contract, Or Principles of Political Right.

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

Sackur, J., 2009. L'introspection en psychologie expérimentale. Revue d'histoire des sciences 62:349.

Sahlins, M., 1972. Stone age economics. Aldine - Atherton, Inc, Chicago and New York.

Sánchez, A. and J. a. Cuesta, 2005. Altruism may arise from individual selection. Journal of theoretical biology 235:233–40.

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

Schäfer, M., D. B. M. Haun, and M. Tomasello, 2015. Fair Is Not Fair Everywhere .

van Schaik, C. P. and P. M. Kappeler, 2006. Cooperation in Primates and Humans: Mechanisms and Evolution. Springer.

Schmidt, M. F. H. and J. a. Sommerville, 2011. Fairness expectations and altruistic sharing in 15-month-old human infants. PLoS ONE 6.

Schokkaert, E. and B. Overlaet, 1989. Moral Intuitions and Economic Models. Social choice and Welfare Pp. 19–31.

Schwagmeyer, P. L., 2014. Partner switching can favour cooperation in a biological market. Journal of Evolutionary Biology 27:1765–1774.

Seyfarth, R. M. and D. L. Cheney, 1988. Empirical tests of reciprocity theory: Problems in assessment. Ethology and Sociobiology 9:181–187.

Sherratt, T. and G. Roberts, 1998. The evolution of generosity and choosiness in cooperative exchanges. Journal of Theoretical Biology Pp. 167–177.

Shweder, R. a., N. C. Much, M. Mahapatra, and L. Park, 1997. The "Big Three" of Morality (Autonomy, Community, Divinity) and the "Big Three" Explanations of Suffering.

Silk, J. B., 2006. Practicing Hamilton's rule: kin selection in primate groups. *in* Cooperation in Primates and Humans: Mechanisms and Evolution.

Silk, J. B., S. F. Brosnan, J. Vonk, J. Henrich, D. J. Povinelli, A. S. Richardson, S. P. Lambeth, J. Mascaro, and S. J. Schapiro, 2005. Chimpanzees are indifferent to the welfare of unrelated group members. Nature 437:1357–1359.

da Silva, R., G. a. Kellermann, and L. C. Lamb, 2009. Statistical fluctuations in population bargaining in the ultimatum game: static and evolutionary aspects. Journal of theoretical biology 258:208–18.

Skitka, L. J., 2012. Cross-Disciplinary Conversations: A Psychological Perspective on Justice Research with Non-human Animals. Social Justice Research 25:327–335.

Skyrms, B., 1996. Evolution of the Social Contract. Cambridge University Press.

Sloane, S., R. Baillargeon, and D. Premack, 2012. Do Infants Have a Sense of Fairness? Psychological Science 23:196–204.

SM, 2015. All supplementary materials for the thesis can be downloaded online at http://stephanedebove.net/?p=208. Ph.D. thesis.

Smith, V. L., 2010. Theory and experiment: What are the questions? Journal of Economic Behavior and Organization 73:3–15.

Sperber, D. and N. Baumard, 2012. Moral Reputation: An Evolutionary and Cognitive Perspective. Mind & Language 27:495–518.

Stahl, I., 1977. An n-person bargaining game in the extensive form. *in* Mathematical Economics and Game Theory, vol. 1.

Stevens, J. R., 2010. Donor payoffs and other-regarding preferences in cotton-top tamarins (Saguinus oedipus). Animal Cognition 13:663–670.

Stevens, J. R. and M. D. Hauser, 2004. Why be nice? Psychological constraints on the evolution of cooperation. Trends in Cognitive Sciences 8:60–65.

Stoop, J., 2013. From the lab to the field: envelopes, dictators and manners. Experimental Economics Pp. 1–10.

Summers, K., 2005. The evolutionary ecology of despotism. Evolution and Human Behavior 26:106–135.

Summerville, A. and C. R. Chartier, 2012. Pseudo-dyadic "interaction" on Amazon's Mechanical Turk. Behavior Research Methods Pp. 116–124.

Sylwester, K. and G. Roberts, 2010. Cooperators benefit through reputation-based partner choice in economic games. Biology letters 6:659–62.

Sylwester, K. and G. Roberts, 2013. Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. Evolution and Human Behavior 34:201–206.

Szolnoki, A., M. Perc, and G. Szabó, 2012. Defense Mechanisms of Empathetic Players in the Spatial Ultimatum Game. Physical Review Letters 109:078701.

Tabibnia, G., A. B. Satpute, and M. D. Lieberman, 2008. The Sunny Side of Fairness. Psychological Science 19:339–347.

Thomson, J. J., 1985. The Trolley Problem. Yale Law Journal 94:1395.

Todd, P. M. and G. Gigerenzer, 2000. Précis of Simple heuristics that make us smart. The Behavioral and brain sciences 23:727–741; discussion 742–780.

Tomasello, M., A. P. A. A. P. Melis, C. Tennie, E. Wyman, and E. Herrmann, 2012. Two Key Steps in the Evolution of Human Cooperation. Current Anthropology 53:673–692.

Tricomi, E., A. Rangel, C. Camerer, and J. O'Doherty, 2010. Neural evidence for inequality-averse social preferences. Nature 463:1089–1091.

Trivers, R., 1971. The evolution of reciprocal altruism. Quarterly review of biology 46:35–57.

Trivers, R., 2006. Reciprocal altruism: 30 years later. *in* Cooperation in Primates and Humans: Mechanisms and Evolution. Springer.

Turiel, E., 2002. The Culture of morality: Social development, context, and conflict.

Van Leeuwen, E. J. C., E. Zimmermann, and M. D. Ross, 2011. Responding to inequities: gorillas try to maintain their competitive advantage during play fights. Biology letters 7:39–42.

Vehrencamp, S. L., 1983. Optimal Degree of Skew in Cooperative Societies. American Zoologist 23:327–335.

Walster, E., E. Berscheid, and G. W. Walster, 1973. New Directions in Equity Research. Advances in Experimental Social Psychology 25:151–176.

Wang, X., X. Chen, and L. Wang, 2014. Random allocation of pies promotes the evolution of fairness in the Ultimatum Game. Scientific reports 4:4534.

Warneken, F., K. Lohse, A. P. Melis, and M. Tomasello, 2011. Young children share the spoils after collaboration. Psychological science 22:267–73.

Warneken, F. and M. Tomasello, 2009. Varieties of altruism in children and chimpanzees. Trends Cogn Sci 13:397–402.

Wascher, C. a. F. and T. Bugnyar, 2013. Behavioral Responses to Inequity in Reward Distribution and Working Effort in Crows and Ravens. PLoS ONE 8.

West, S. a. and A. Gardner, 2010. Altruism, spite, and greenbeards. Science (New York, N.Y.) 327:1341–4.

West, S. a., A. S. Griffin, and A. Gardner, 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of evolutionary biology 20:415–432.

West, S. S. a., C. E. Mouden, A. Gardner, and C. El Mouden, 2011. Sixteen common misconceptions about the evolution of cooperation in humans. Evolution and Human Behavior 32:231–262.

Wiessner, P., 1996. Leveling the hunter: constraints on the status quest in foraging societies. Pp. 171–192, *in* Food and the Status Quest: An Interdisciplinary Perspective, p. wiessne ed. Berghahn Books, Oxford, UK.

Wilensky, U., 1999. NetLogo.

Winking, J. and N. Mizer, 2013. Natural-field dictator game shows no altruistic giving. Evolution and Human Behavior 34:288–293.

Woodburn, J., 1982. Egalitarian Societies. Man 17:431–451.

Zizzo, D. J., 2011. Do Dictator Games Measure Altruism ? Handbook on the Economics of Philanthropy, Reciprocity and Social Enterprise .

Zollman, K. J., 2008. Explaining fairness in complex environments. Politics, Philosophy & Economics 7:81–97.

Abstract

Humans care about fairness and are ready to suffer financial losses for the sake of it. The existence of such costly preferences for fairness constitutes an evolutionary puzzle. Recently, some authors have argued that human fairness can be understood as a psychological adaptation evolved to solve the problem of sharing the costs and benefits of cooperation. When people can choose with whom they want to cooperate, sharing the costs and benefits in an impartial way helps to be chosen as a partner and brings direct fitness benefits. In this theory, partner choice is thus the central mechanism allowing the evolution of fairness. Here, we offer an interdisciplinary study of fairness to put this theory to the test. After a review of competing theories (Paper 1, in review), we build game-theoretical models and agent-based simulations to investigate whether partner choice can explain two key aspects of human fairness: the wrongness to take advantage of one's strength to exploit weaker people (Paper 2, *Evolution*), and the appeal of distributions where the reward is proportional to the contribution (Paper 3, in review). We show that partner choice succeeds at explaining these two characteristics. We also go towards more realistic and mechanism-oriented simulations by trying to evolve fair robots controlled by simple neural networks. We then test the theory empirically, and show that partner choice creates fairness in a behavioral experiment (Paper 4, *Proceedings of the Royal Society B*). We develop a collaborative video game to assess the cross-cultural variation of fairness in distributive situations, and present results coming from a Western sample (Paper 5, in preparation). We review the experiments looking for fairness in non-human animals, and discuss why fairness would have been more prone to evolve in humans than in any other species, despite partner choice being an evolutionary mechanism far from restricted to the human species. Finally, we discuss three common misunderstandings about the partner choice theory and identify interesting directions for future research.

Keywords : human fairness, equity, justice, partner choice, biological markets, cooperation, morality.