On Altruism: a Behavioral Study

by

Carolina Di Tella

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

September 2020

Wei Ji Ma

If I am not for myself, who will be for me?

If I am only for myself, what am I?

Rabbi Hillel

Dedication

To my father, who always encouraged me to follow my dreams.

Acknowledgements

There are no words to describe the gratitude I feel towards my two advisors, Paul Glimcher a Wei Ji Ma. The word 'eternal' comes to mind. As I stand at the end of this stage, I look back and I see dedicated teaching, inspiring passion for the process of inquiry, and more nurturing love that I could have ever imagined.

Paul was the first to give me a chance. The journey of transitioning from economics to neuroscience was long and arduous. It could very well be described as solo whitewater rafting and almost drowning. I finally washed ashore at an SfN conference, where I met Paul's lab member, Kenway Louie. We discussed science and dreams. He said, "write to Paul asking for a meeting, and copy me". Paul gave me an interview and inadvertently opened up a world of opportunities for me.

Wei Ji came later into my life but had no secondary role. We started working together on my first rotation, and we never stopped. Always kind, always understanding, I would consider myself successful if a little bit of his ways stuck on me. I know I'll be striving to follow his example until the day I die.

The incredible excitement about pursuing the PhD in neuroscience at NYU was unfortunately marred by a string of three deaths of special people in my life in less than three years. Both Paul and Wei Ji were incredibly supportive when life couldn't throw any more dead bodies at me. I couldn't have done it without them.

I also want to acknowledge my late father, Torcuato, who always encouraged me to follow my dreams. When in doubt, he always said, "buy the books", and that made a huge difference. He gave me total freedom to carve my own path, sometimes to a dangerous degree, but I wouldn't have had it any other way.

With these three giants duly recognized, I want to acknowledge people who made a huge difference during the transition to neuroscience and during my wonderful years at NYU. First, I wouldn't have made it to the shore without the advice of Mariano Sigman, E.J.Chichilnisky, and Diego Gutnisky. Once at NYU, my cohort classmates quickly became like a family: Andra, Christina, Kristina, Emily, Oliver, Evan, and Alex; thank you for making it magical. Paul's lab members became like a second family right away: Silvie, Mel, Hui-Kuan, Sam, Kenway, Jan, David; I couldn't have asked for a better lab environment. Upon my return to NYC after a much-needed break from death and destruction, Wei Ji's lab members welcomed me with the friendly attitude that characterizes the lab; thank you guys for that. I also have to thank my aunt, Graciela, who happens to live in NYC and gave me unconditional support when I needed it most. I will never forget it, and I hope I was able to do the same for you. Back at home in Argentina, my friends and family were always with me in spirit, giving me the strength to venture out. I need to especially mention my brother Sebastian, my mom Tamara, and my friends Irene and Belen, who was the first to go. Today, I have to thank Tuvia for restoring my peace.

I would also like to thank my committee members David Heeger and Mike Hawken, who at the same time challenged me and supported me, and my external reader Andrew Schotter, who kindly accepted my request to read this thesis and offered valuable advice on the last chapter. In more general terms, the entire NYU atmosphere was just wonderful, from staff to professors and students from other cohorts. Finally, none of this would have been possible without the generous funding from the NYU MacCracken Fellowship.

Preface

Chapters 2 and 3 are being prepared for resubmission. These chapters describe work done together with Wei Ji Ma and Paul Glimcher.

Chapters 4 and 5 describe work done with Wei Ji Ma. Chapter 4 may be included in the aforementioned manuscript.

Abstract

Altruism, the tendency to value others, is at the center of how we interact with each other. It has relevance both at the individual and the societal level. In this thesis, I explore a series of questions about human altruism. How altruistic are we? Is altruism the product of self-control over selfish impulses or, on the contrary, is it an impulse itself? And, if we are indeed naturally altruistic, how is it possible, from an evolutionary point of view?

These questions have a long tradition of thought behind them. Here, I attempt to contribute to the discussion by supplementing a neuroeconomics approach with a novel 'maximum-information' trial selection method, which I developed together with my advisors. I call the final product the 'DGM' trials after the initials of the people involved. While designed specifically for our task, the same trial selection procedure can be applied to other tasks and to address different questions. I present this methodological contribution first. I then present a series of experimental attempts to answer the questions posed in the opening, while making use of the DGM trials.

After a brief introduction to the subject matter in chapter 1, in chapter 2 I describe in detail the DGM trial selection method for a two-alternative forced-choice (2AFC) Dictator Game (DG) task. Initiated prior to the first experimental run, I developed this method iteratively, incorporating the data obtained in earlier stages as priors on the following ones. This provided improved trials as I progressed on the experimental projects. The method builds on the concept of mutual information. Mutual information between two variables is a measure of how much knowing one of the variable reveals about the other. I selected trials for the DG task to maximize the mutual information between the observable and unobservable variables of interest, as estimated using simulations. By revealing the most amount of

information, the selected trials maximized the power of my statistics, which allowed me to perform all of the analyses at the individual level, without having to rely on group averages.

In the third chapter, I present the first experimental project, where I performed a systematic model comparison of other-regarding preferences using the DG task and the first set of DGM trials. Across this and the following experiments, the 'Other' was either a friend of the decision maker, or a stranger, or both. I found that the Charness-Rabin model outperforms the Rawlsian and Cobb-Douglas models. This rejects extreme egalitarianism but recognizes status-dependent behavior. The Charness-Rabin model also outperforms the status-dependent Fehr-Schmidt model of inequality aversion. This model describes preferences that are pro-social when the decision maker is better off, but antisocial when they are worse off. In contrast, the Charness-Rabin model allows for a socially efficient inclination that is less giving when the decision maker is worse off, but never crosses over to the anti-social domain. Most people fall in this category. Interestingly, in our first sample, the vast majority of subjects showed the same disposition towards friends and strangers.

In the fourth chapter, I address the question of whether altruism is an impulse or rather a product of self-control. To do so, I added a N-Back working memory task in simultaneous to the DG decision-making process. This cognitive-load task drains the decision makers of resources needed to exert self-control, shall they need it. The classic view of altruism as a social demand that we strive to conform to predicts that cognitive load will reduce pro-sociality. The alternative hypothesis of altruism as an impulse predicts that pro-sociality will remain the same or increase. I hypothesized, leveraging on the potential for individual level analysis provided by the DGM trials, that there would be a great deal of individual variability. Indeed, I found that the effect of cognitive load goes in different directions for different individuals. Interestingly, in half of subjects that showed an effect of

cognitive load on other-regarding preferences towards friends and strangers, its effect tended to go in the same direction with both partners.

In the fifth chapter, I investigate the evolutionary origins and viability of altruism. I tested the group-selection theory, which posits that altruistic groups outperform selfish ones. I used a nested Patent Race through Public Goods Game (PGG), where two groups competed for a prize via voluntary contributions. The two groups differed in the level of altruism of its members, who were screened beforehand using the DG task. We found that, indeed, altruistic groups tended to outperform less altruistic ones, but the effect size, when significant, was small. This could be partly explained by the specific configuration of experimental parameters, such as cost and benefit of cooperating, and so, other incentive structures are recommended for future research on this design.

I finalize with a conclusions chapter, were I summarize the work undertaken in this thesis, contextualize it in relation to the broader field, and point out improvements for future work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Why Study Altruism

The question of altruism is at the core of every philosophy that aims to achieve an egalitarian or just society. Even if not stated explicitly among their tenets, without it the push for equality cannot be achieved peacefully, it is not sustainable, and is not compatible with freedom. Without peace and freedom, equality is nothing more than an appearance, a shallow rallying cry, a tool to sustain and perhaps even deepen inequality and domination.

The extent to which humans are naturally inclined to consider others in their decisions and make sacrifices for them is paramount at the moment of evaluating which systems are possible and even desirable. For instance, absolute material equality may require such a degree of self-negation -specially in the face of scarcity- that only by denying our very nature can we conceive of this goal.

Moreover, generosity is likely to be moderated by the relationship we have with those who receive it, and so, what works in small and tight communities may not work in mass societies, where relationships become more distant and abstract. Whether it is in-group versus out-group, kin versus non-kin, or friend and foe, these natural inclinations must be considered in any system that has any chance at succeeding.

With these considerations in mind, we asked a series of questions about altruism in humans. How altruistic are we? Are we really ever altruistic? Or are we trying too hard? And, if we are altruistic, how is that even possible?

1.2 Ecological Validity – Some Considerations

Measuring altruism -the extent to which we consider others as ends in and of themselves- poses certain unique challenges in terms of ecological validity. In real life, altruism is expressed in all sorts of currencies: time dedicated to someone else's projects, risks to our own life or limb to save another's, reputational damage taken to protect someone else, etc. All of these ways in which altruism is expressed in real life constitute the heart of what makes it so important to us. However, expressed in those very real currencies, it is almost impossible to measure.

Researchers interested in a quantitative understanding of altruism have had to content themselves with monetary decision-making experiments in laboratory settings. The drawback of limiting the study of altruism to the monetary, laboratory-compatible domain is compensated by the quantitative precision obtained within that limited domain. This sacrifice in 'ecological validity' is a trade-off that cannot be escaped.

Another challenge arises when we need to distinguish altruism from reciprocity. Reciprocity entails pro-social behaviors for motivations that may not be altruistic in nature. To tease altruism and reciprocity apart, we must eliminate the possibility of repayment, of benefits ripped in the future for our generous deeds of today. In the language of game theory, we need to eliminate the effect of ‘repeated interactions’.

It is difficult to conceive of the total elimination of this variable in real life. Even when interacting with strangers that we do not expect to see again, the possibility always remains. In laboratory experiments, we can easily eliminate the scope of repeated interactions by both keeping anonymity and limiting the information about choices that we divulge to those affected. If someone cannot keep track of the favors given or received, one cannot repay or demand repayment. But this also has its limitations. The fact that there is no rational, logical expectation of seeing someone again does not mean that we can rule out that, in the back of our subjects’ minds, the possibility always remains.. This caveat, also, cannot be escaped.

1.3 The Classic Understanding

The classic understanding of human nature until rather recently was that we are selfish. Only culture and social norms were thought to provide a compensatory motivation for overcoming our selfishness and be considerate of others. The Homoeconomicus -a ‘rational’ man who thought only of himself and in material terms- reigned the day.

Under this view, we engage in prosocial behavior either to reap future benefits, to maintain a reputation that will assist in that process, or to strategically manipulate others for our advantage. Certainly not out of pure consideration of others. From biology to economics, selfishness was the default assumption, and the onus was on researchers to prove the existence of altruism -not the other way around.

Even in a situation where there is no potential gain, whatsoever, in the horizon, from an apparently altruistic act, skeptic etiquette demanded an explanation such as the theory of ‘Warm Glow’ (Andreoni, 1990). ‘Warm Glow’ is the idea that we do good things because it makes us feel good about ourselves –but, crucially, not out of valuing others. It is possible to conceive of a person mindlessly ‘chasing’ the ‘rush of feeling good about themselves’ and that is, understandably, not altruism. However, just feeling good after doing something that one finds meaningful should not be a disqualifier, it should go without saying. Feeling ‘warm’ does not a pure intention disqualify.

While there is an element of truth in the theory of warm glow, it should be noted that the extreme ‘warm glower’ would, for instance, make a donation to a philanthropic organization and not care whether the money 1-actually had a positive impact, 2- was burnt on a bonfire, 3- or was embezzled by the organization’s leaders. Nor would they care which cause they are donating to, as long as ‘a’ donation is being made. It is unlikely that someone would behave as a perfect ‘warm glower’. Warm glow is not enough, there must be something else that motivates altruistic behavior.

1.4 A Neuroeconomics Framework

We approach this exploration of altruism using a neuroeconomics framework. Neuroeconomics combines principles of economics and neuroscience (Glimcher, 2011). In its simplest terms, the neuroscientific view relies on the understanding that human beings are but one species amongst many in the animal kingdom. It also relies on the idea that the mind (or psyche) is nothing more than a manifestation of brain activity. The economic view, on the other hand, relies on the idea that decision-makers are, for the most part, rational agents trying to satisfy their needs and wants the best way they can. It also relies on the equivalent idea that fairly rational agents can be understood as maximizing a 'utility function' –albeit admittedly, with noise. Both conceptions are compatible and nurture each other in interesting ways.

1.4.1 Neuroeconomics I: the Neuroscience in Neuroeconomics

The neuroscience of decision-making is broadly divided into two categories: perceptual and value-based. Perceptual decision-making is in regard to the external world: there is a certain objective truth that we are trying to know through our senses. An evolutionary relevant example may be the need to determine whether there is a lion lurking in the grass, or is it just a figment of our imagination? There is a right and a wrong answer. The lion is either present or not. And the consequences of getting it wrong are dire. We evolved to solve this perceptual problem rather efficiently, without expending unnecessary energy on erratic behavior. (Bisley and Goldberg, 2010; Gold and Shadlen, 2007)

Value-based decision-making is in regard to our internal idiosyncratic preferences - our values. Would I like my dessert tonight to be tiramisù or crème brûlée? In value-based decision-making, there are no right or wrong answers, only preferences. But there is such a thing as choosing the most preferred option. Presumably, the same system that evolved to help us perceive the external world fairly correctly also helps us orient ourselves towards what we prefer and value more. (Glimcher, 2014; Sugrue et al., 2005)

While the two types of decisions differ greatly in the nature of what the choice is about, they share the same neural mechanisms for evaluating options and implementing a choice. In both cases there is a system that produces an ordering of options and chooses the one at the top. In perceptual decision-making, the top option will be the most likely explanation of the world; in value-based decision-making, it will be the most preferred good or course of action.

Such is the overlap between perceptual and value-based decision-making that a perceptual problem can easily be formulated in value-based terms. Think of the value of correctly identifying the presence of a lion in the grass, and the cost of sprinting unnecessarily when the lion is not there.

In both perceptual and value-based decision-making, choice can be understood as being guided by a 'decision function' that represents the decision-maker's objectives. In perceptual decision-making, a decision function would take the sensory information that we receive through our senses and produce an ordering of likelihoods of possible explanations for the event in question. It would, for instance, take all of my sensory input and determine whether it is more likely that the lion is present or not. In as much as it is desirable to have a correct understanding of the world, the perceptual decision-making system will tend to select the explanation that is most likely correct.

In value-based decision-making, the decision function is one of subjective-value. It can also be called a 'utility function', a term borrowed from economic theory to represent what our hearts desire. A subjective value or utility function is a mathematical representation of our preferences. It maps input –goods or actions- to values, producing an ordering of options that are more or less desirable to us. In as much as it is desirable to satisfy our hearts' longings, the value-based decision-making system will tend to select the option that is most valuable to us.

1.4.2 Neuroeconomics II: the Economics in Neuroeconomics

The economic framework is remarkably compatible with decision neuroscience. From an economics point of view, people, or individuals of any kind for that matter, are understood to make decisions by ‘maximizing’ a ‘utility function’. This principled process of maximization explains, or ‘rationalizes’, the choices we observe. The economic theory of decision-making poses unique challenges, but also brings enormous value from an epistemological point of view. The greatest challenge, and that which generated most controversy, is the idea of the ‘rational agent’.

Here it might be worthwhile pausing for a moment to discuss rationality. What does ‘rationality’ mean? Let us start by distinguishing the colloquial sense of the word ‘rational’ from its technical sense. In colloquial parlance, ‘rational’ is used as opposite to ‘erratic’, but also as opposite to ‘the passions’, or equivalent to ‘deliberative’. However, this does not make justice to most of our behaviors. Take for instance any of the drive states, such as hunger and thirst. These are hardly elevated cognitive functions. However, no one would argue that when we are hungry it is irrational to focus our attention and direct our behavior towards obtaining food. The hungrier we are, the more focused on food we will be, in disregard of other interests such as getting tickets for a concert of our favorite artist. Emotions are not quite the same as drive states, but a similar argument can be made. Both drive states and emotions (“the passions”) serve the purpose of grabbing our attention and motivating us to engage in specific behaviors that have served us in our evolutionary history (LeDoux and Damasio, 2013; Loewenstein, 2000; Panksepp, 2004; Tooby and Cosmides, 2008). Is it irrational to sprint when we think the lion is there and we are overtaken by fear? Is it irrational to want to punish someone who has offended us and made us angry? (Sell et al., 2009) In this sense, only the opposite of ‘erratic’ is a valid colloquial meaning of the word ‘rational’.

The technical sense of the word is more interesting and serves the epistemological purpose that we briefly mentioned when we introduced this subject. Any understanding of choice that goes beyond the descriptive and attempts to provide a normative framework will inevitably build on a concept like a ‘preference’. The problem with preferences is that they are not observable, and as such, they pose a serious challenge from the scientific point of view. How can we even know that preferences are represented by utility functions, and that people make choices by maximizing them, if we could never directly observe them? And to dare make predictions about choices based on that fantasy concept?

In what follows, we will attempt to make the briefest possible recount of developments in economic theory that culminate in the current understanding of choice and rationality. The first step towards a scientific theory of choice came in 1906 with Wilfredo Pareto’s ’Ordinal Revolution’ (Pareto, 1906). He recognized not only that utility was unobservable, as we have just mentioned, but also that there was a multiplicity of utility functions that could rationalize any set of choices. Since preferences are ultimately about an ordering of desirability, the exact amount of utility ascribed to any option does not have a special meaning. Multiply the utilities of every good by two, and you have the same ordering. Divide them by three, and the ordering remains the same. In fact, any monotonic transformation of a proposed utility function is as valid as the original one. Pareto’s insight established that utility functions need only be ordinal - not cardinal- to serve as a foundation for any scientific theory of choice. This liberated early economists from finding the precise form of the mythical utility function, and from the problem of the incommensurability of utilities between people.

The un-observability bridge began to be crossed in 1938 when Paul Samuelson proposed the theory of 'Revealed Preference' (Samuelson, 1938). Rather than assuming that individuals choose by maximizing a utility function, he focused on how we would go about testing that hypothesis. Choice being the only observable variable here, he asked: what characteristics would a choice set consistent with utility maximization display? If the choice set fails to show those characteristics, it could not have been produced by the maximization of a utility function. If it does display those characteristics, then we cannot rule out that it was indeed produced by such a process. In the latter case, we can then use the theory of utility maximization as a tool to rationalize –explain- those choices. His genius was to provide a test of rationality by clearly stating the conditions where rationality is violated.

In practical terms, Samuelson's rationality test is as simple as this: if I choose a when b is also available, then I either prefer a to b or I am indifferent between the two, but I certainly do not prefer b to a, or I would have chosen it. In a concrete example, if I chose tiramisù over crème brûlée when crème brûlée is also available, then I may love both desserts, but I definitely do not prefer crème brûlée over tiramisù - at least not tonight. This is called the 'Weak Axiom of Revealed Preference' (WARP), whereby I reveal my preferences through the choices I make. Consistency with WARP is a necessary condition for utility maximization. Fail the WARP test, and there is no utility function that could rationalize the observed choices.

In 1950, Houthacker extended the WARP theorem to the 'Generalized Axiom of Revealed Preferences' (GARP) (Houthakker, 1950). GARP provides both necessary and sufficient conditions for rationality. It simply applies the same logic of revealed preference to longer strings of objects: if I prefer a to b, and b to c, then I must also prefer a to c. This is called transitivity and is the sole technical requirement of rationality. In other words, just don't be erratic!

The final step in the evolution of modern economics as we find it today is the theory of Expected Utility. It goes above and beyond elemental rationality to allow for a more sophisticated utility representation that incorporates uncertainty into the equation. Proposed by von Neumann and Morgenstern in 1944, it adds three additional conditions to the basic one of transitivity (Morgenstern and Von Neumann, 1953). These are very technical, mathematical requirements that allow for the existence of a utility function that comes from multiplying utilities by their probabilities. Take, for example, the utility of briefly parking by a fire hydrant to get something from the store. The calculation will consider the probability of being caught by an officer and the amount of the fine. Just to mention them here, the additional conditions are: i-completeness, ii- continuity, and iii- independence of irrelevant options. This last axiom is where most of the violations of rationality are observed, as is the case with the decoy effect.

Because there is only one way of being rational, but an infinite number of ways of being irrational, rationality is the only connection between unobservable preferences and observable choices. Revealed preference is actually central in our everyday processing and understanding of the external world, especially when making inferences in social contexts. For instance, if someone keeps having love affairs, we can safely infer that this person is not committed to a monogamous relationship with their partner. We can make these inferences only by assuming a basic level of rationality (Baker et al., 2008; Ullman et al., 2009).

While we seem to violate rationality in a fairly systematic way, those failures seem to be better understood as biological constraints on our computational machinery more than a negation of the principle of rationality per se (Louie et al., 2013). Between 'perfectly irrational' and 'imperfectly rational' beings, we lean on the rational side of things. Even non-human animals display exquisitely rational behavior, both in the wild and in laboratory experiments (Glimcher et al., 2005). Neuroeconomics has brought all of these decision aspects together in an incredibly parsimonious fashion.

1.5 Social Preferences

As we have mentioned, the original utility functions were utterly selfish. The variables that went into it, material and self-regarding. This was understandable in the early days of the discipline, when it was just beginning to lay out its principles. At that time, it was called 'political economy' and its practitioners were primarily concerned with pecuniary objects: loans, interest rates, production, consumption, work, etc. It was not the time to incorporate non-pecuniary objects such as moral concerns or any other abstract value. This gave Homoeconomicus -the idea of the rational man- a bad rep. As experimental data came in showing people to be more complex than that (Allais, 1953), it weakened its credibility as well.

In the realm of social decision making, positive offers in both the ultimatum and the dictator game were first interpreted as irrational (Forsythe et al., 1994; Güth et al., 1982), but these findings later motivated researchers to think that we needn't be completely selfish after all. In the 90's, several researchers proposed models of utility functions that take into consideration the welfare of others, and even consider abstract moral values such as fairness. This began a road towards a more complex, perhaps even sentimental, Homoeconomicus; a rational decision maker that cares for something beyond his own material gain. Rabin (1993) developed the concept of fairness equilibrium, with agents that are inclined to help those who help them and hurt those who hurt them. Levine (1998) proposed a similar model, where agents vary in their degree of altruism and spitefulness towards others. Perhaps the most influential model was proposed in 1999 by Fehr and Schmidt. Inspired by the economics' symbolic universe, the Fehr-Schmidt model introduced others through the inequality aversion motive. This model incorporated the use of relative status-dependent preferences, where the weight for the other depends on whether the decision maker is better or worse off than the recipient. Crucially, difference aversion requires negative weights for others when the decision maker is worse off, which implies antisocial dispositions. Charness-Rabin (2002) improved on this status-dependent model by freeing it of the difference aversion constraint. Cox and Sadiraj (2005) proposed the egocentric altruism model, which applies the status-dependent principle to a non-linear constant elasticity of substitution (CES) function. Bolton and Ockenfels (2000) proposed the ERC model, which stands for 'Equity, Reciprocity, and Competition'. Most of them carried on limited testing, with varying degrees of success. No one won, but Fehr-Schmidt took the prize anyway.

At one point, with so many models representing different ways of considering others, a nomenclature clarification became necessary. Some started to distinguish social preferences from distributional preferences, in reference to how people would like resources to be distributed among others, once they themselves are out of the equation (Fisman et al., 2007). Another distinction is usually applied for purely non-strategic scenarios. Sometimes, there are higher order social concerns that might influence our expression of altruism, such 'saving face', when someone would cooperate but does not out of fear the other will not. To distinguish social preferences in these strategic scenarios from non-strategic ones, social preferences in the latter context are referred to as other-regarding preferences. These are what we will be focusing on.

The wide variety of other-regarding utility functions suggest different ways of considering others, and therefore can shed light on the nature of altruism. I will develop this issue in more detail in chapter 3 - A Systematic Characterization of Other-Regarding Preferences. For now, I will just mention that an experiment conducted by Andreoni and Miller (2002), proved that regardless of the functional form, altruism can be, and typically is, rational.

1.6 Structure of the Thesis

Going back to the questions that guided this thesis, I asked: how altruistic are we? Are we really ever altruistic? Are we trying too hard? And, if we are indeed altruistic, how is that even possible? I explored these questions through three avenues, which will correspond to the three experimental chapters of this thesis, chapters 3 to 5. In the first experimental chapter, chapter 3 - A Systematic Characterization of Other-Regarding Preferences, I attempted to measure a baseline regard for others of differing degrees of closeness (friends and strangers), in a neutral context. I investigated, for instance, whether people behaved differently when they were better off or worse off. And if so, how different and in what way? This entailed an exercise in model comparison, where each candidate model represented a different kind of other-regarding preference.

To aid me in that endeavor, I developed a novel trial selection method designed to maximize the power of our statistics. Based on information theory, we developed this method in incremental steps, first to identify the best performing model of other-regarding preferences and then to maximize the precision of the parameter estimates at the individual level. This efficient trial selection method provided us with a solid foundation for subsequent work. I present this methodological work first, in chapter 2 - An Information Efficient Trial Selection Method.

In the second experimental chapter, chapter 4 - The Effect of Cognitive Load on Other-Regarding Preferences, I introduced cognitive load in simultaneous with the decision-making task. Cognitive load tasks drain subjects of mental resources that would otherwise be used for impulse control. This allowed me to query to what extent expressions of regard for others are the result of natural impulses as opposed to a desire to conform to social demands. Behaviors that arise from social demands are arguably weaker than natural dispositions or drives. Systems that rely on weaker forces will inevitably be weaker compared to those that rely on natural, stronger forces. Furthermore, systems that negate strong natural forces are doomed to implode.

In the third experimental chapter, chapter 5 - A Test of the Group Selection Theory of Altruism, I considered an evolutionary hypothesis for the existence of altruism towards strangers as a natural disposition. Altruism towards non-kin has been a longstanding challenge to evolutionary biologists. Most acts of altruism entail a cost, and individuals that make sacrifices for others lose in relative fitness, making it less likely that their genes will pass on to the next generation. Lacking an opposing driving force to favor the genes that produce altruistic behavior, in time they will disappear from the population.

The theory of group-selection provides a rationale for this compensatory driving force. Group-selection theory rises one level up, from individuals to groups, including groups of the same species. Concretely, group selection could favor a certain level of altruism in the population if groups with altruist members outperform (grow faster than) groups composed solely of selfish individuals. We tested an adapted version of this hypothesis for a modern human context, not through reproduction but through financial gains, in a nested game of ‘Patent Race through Public Goods Game’.

Finally, I close this thesis with a conclusion chapter that looks back at the set of experiments and interprets their results in light of a bigger picture, suggesting improvements for future work.

Chapter 2

An Information Efficient Trial Selection Method

2.1 Introduction

The study of altruism using a computational approach is relatively recent. The classic, selfish and materialistic, Homoeconomicus started acquiring a beating heart in the 1990's, with a series of models of social preferences being proposed from the halls of Economics departments (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Cox and Sadiraj, 2005; Levine, 1998). The most influential was proposed by Fehr and Schmidt in 1999 (Fehr and Schmidt, 1999). This seminal model incorporated others into the equation via the inequality-aversion motive.

The testing of these models also has a relatively short history. It starts with the original Dictator Game (DG), where a fixed amount of money is divided unilaterally by "the dictator" between himself and another person (Forsythe et al., 1994). Later on, testing evolves to menus of a handful of options (Cox and Sadiraj, 2005; Engelmann and Strobel, 2004), to several options (Charness and Rabin, 2002), and finally, to the (infinite-options) budget-set experiments, linear (Andreoni and Miller, 2002) and step-wise (Fisman et al., 2007).

Linear budget sets, first used by Andreoni and Miller 2002, expanded the dictator game in that i- they are continuous, and ii- they offer different prices (or costs) of giving (see Figure 2.1(a)). This second attribute is interesting because it allows for the testing of rationality (Andreoni and Miller, 2002). It is limited however in the fact that 'walrasian' budget sets (budget sets with positive prices) cannot distinguish, for instance, between 'narrowly selfish' and competitive individuals. Narrowly selfish individuals simply disregard others. Others do not enter into their utility function. They would not make sacrifices for others, but neither would they avoid benefiting them[1]. Competitive agents, on the other hand, are willing to incur in costs to put others down.

Stepwise budget sets, first proposed by Fisman et al, 2007, break out of the 'walrasian' limitation by offering pieces of budget sets where giving to the other is free (the price or cost of giving is zero) (see Figure 2.1(b)). This allowed them to distinguish between participants who, given the 'free' option, systematically chose the prosocial extreme of the budget set (lexicographic preferences), or the antisocial one (competitive), or chose randomly along the two ends (narrowly selfish).

We noticed that, while the methods for measuring other-regarding preferences had evolved monumentally, there was still an ad-hoc element, at least in establishing the parameters of the budget sets. We set out to contribute to this effort by proposing a systematic method of trial selection that is guided by the maximization of mutual information

[1] If they are lexicographic, they would gladly benefit others but only if it came at no cost

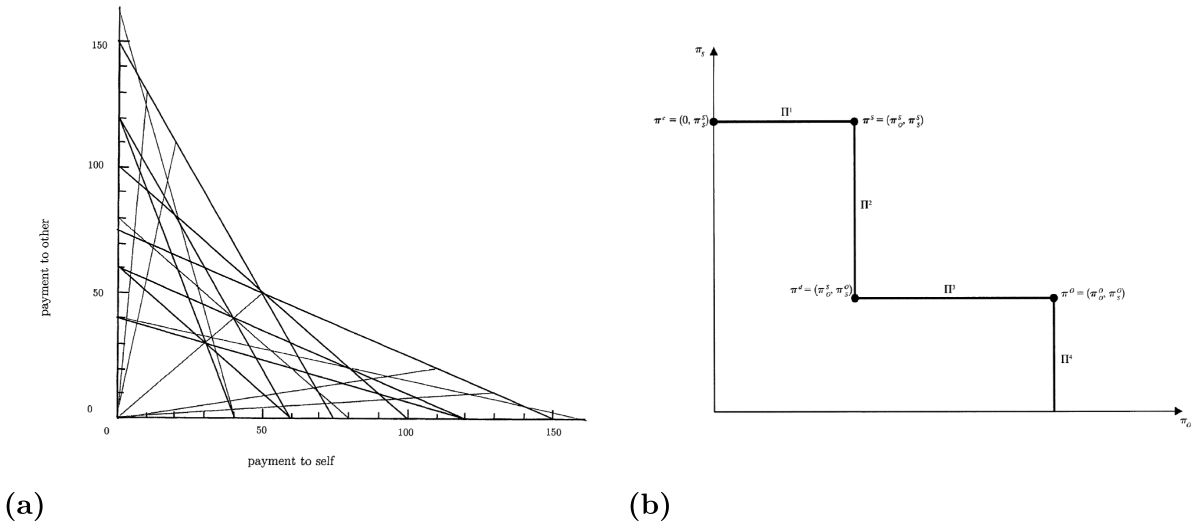

Figure 2.1 (a) Linear Budget Sets. Reproduced from Andreoni and Miller 2002. **(b)** Step-wise Budget Sets. Reproduced from Fisman, Kariv and Markovits 2007

between data and models.

2.2 Mutual Information

Mutual information between two variables is a measure of their mutual dependence. It indicates to what extent knowing one of the variables reduces uncertainty about the other. Take for example the categorical variables M (for models of interest), and r (for the possible responses to a specific 2AFC dictator trial). The mutual information between the two variables is given by the following equation:

$$MI\left(r,M\right)=\sum_{M}\sum_{r}p\left(r,M\right)log\frac{p\left(r,M\right)}{p\left(r\right)p\left(M\right)}$$

Notice that, if r and M were independent, the joint probability of r and M, $p(r,M)$, would be equal to the product of the probabilities of r and M separately, $p(r)p(M)$, making the ratio inside the logarithm equal to 1, and the entire equation equal to zero. In that case, there would be no mutual information between the two variables, and knowing about one of them would tell us nothing about the other. If, on the other hand, r and M were related, then observing one of the variables would tell us something about the other.

The question is then: how much about the models can I know from observing a response to a specific trial? Or put in statistical terms, how much uncertainty about the models is reduced by observing such a response? By our method, each trial can be ranked in terms of the amount of information they provide, information in the sense of reducing uncertainty.

The objective was to eventually rank the trials from most to least informative, and select the most informative ones to use in the experimental sessions. To do this, we had to rely on simulations. We used simulations to compute, separately for each candidate trial, the mutual information between the candidate models and the choices produced by those models.

2.3 Simulations

We considered a list of candidate models and a list of candidate trials. The method of trial selection that we propose here is, per se, independent of the models and trials being compared. Other, better curated lists of models and trials can be evaluated using this method.

2.3.1 Candidate Trials

We considered approximately 3000 candidate trials. Trials consisted of a 2AFC dictator task. On each trial, there were two options. Each option consisted of an amount of money for the decision-maker and an amount of money for another person. Possible payoffs to Self and Other ranged between \$3 and \$100. We created reference options by taking six steps of \$15 on each dimension, from \$15 to \$90. The 36 possible combinations ranged from (\$15, \$15) up to (\$90, \$90). For each of the 36 reference options, we created 64 alternative options around it, in one of 16 possible angles and one of 4 possible radii. We then created additional trials by taking all unique combinations of reference points. This yielded a total of 2934 candidate trials (See Figure 2.2(c)).

By our method, we aim to rank these candidate trials by the amount of information they provide about the potential models of other-regarding preferences.

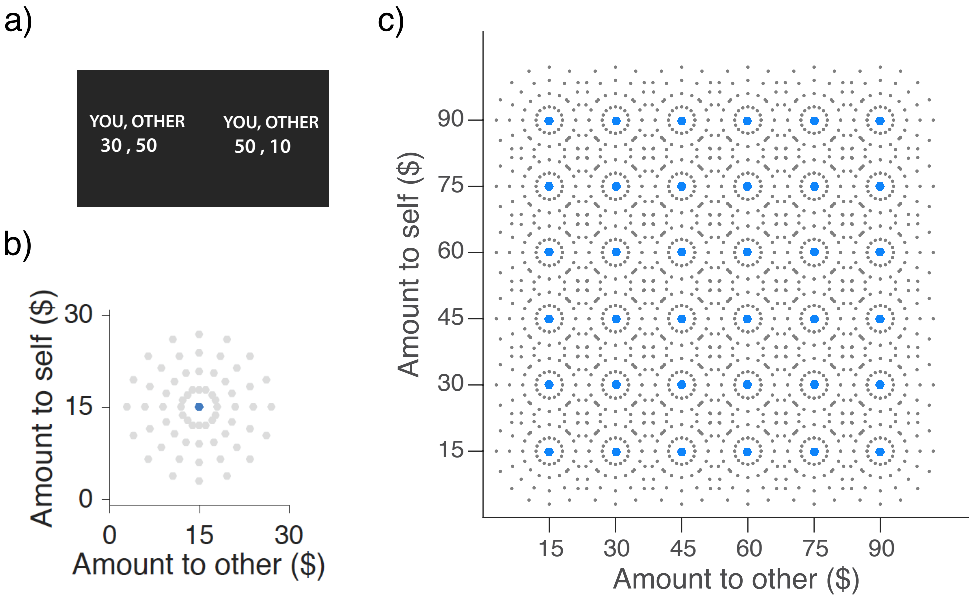

Figure 2.2 Candidate Trials. **(a)** One trial of a 2AFC dictator game as it would be presented to real subjects. Each trial consists of 2 options; each options consists of 2 amounts, one for self and one for other **(b)** Close up of an example reference point (blue) and 64 alternatives around it (grey). Alternatives are given by 16 possible angles and 4 possible radii. **(c)** All candidate trials. There are 36 reference points, with 64 alternatives around each one, and every combination of reference points. This yields a total of 2934 candidate trials. A purple line connects the the 2 options of the 100 'most informative' trials selected for experiment 1.

2.3.2 Candidate Models of Other-Regarding Preferences

We considered a list of candidate models that we were interested in evaluating. We state here only their functional form, for more on their meaning in terms of other-regarding preferences and why we chose them, see chapter 3 - A Systematic Characterization of Other-Regarding Preferences, section 3.2.5 - Models of Other-Regarding Preferences).

We settled on testing the Charness-Rabin model, a Cobb-Douglas social model, and a modified Rawlsian model. Others may very well be considered.

Charnes-Rabin model:

$$U\left(x_{self}, x_{other}\right) = \begin{cases} \left(1-\rho\right) x_{self} + \rho x_{other}, & \text{if } x_{self} \geq x_{other} \\ \left(1-\sigma\right) x_{self} + \sigma x_{other}, & \text{if } x_{self} < x_{other} \end{cases}$$

$$\rho, \sigma \in [-1, 1]$$

Cobb-Douglas model:

$$U\left(x_{self}, x_{other}\right) = {x_{self}}^{\gamma} {x_{other}}^{(1-\gamma)}$$

$$\gamma \in [0, 1]$$

Rawlsian model:

$$U\left(x_{self}, x_{other}\right) = \min\left(\omega x_{self}, x_{other}\right)$$

$$\omega \in [0,1]$$

We considered 1000+ parameter combinations per model, which we call ‘agents’ (specified below, in section 2.3.3). To numerically approximate these probabilities, we simulated responses from each model, assuming a linear grid of parameter combinations within each model, which we will call agents. For the Rawlsian model, the grid consisted of 35 steps for β (between 1 and 10) and 35 steps for ω (between 0 and 1). For the Cobb-Douglas model, the grid consisted of 35 steps for β (between 1 and 10) and 35 steps for γ (between 0 and 1). For the Charness-Rabin model, the grid consisted of 10 steps for β (between 1 and 10), 11 steps for ρ (between -1 and 1), and 11 steps for σ (between -1 and 1). To account for decision noise, each simulated agent responded 10 times to the candidate trials.

2.3.3 Mapping Utility to Choice

To model an agent’s choice between two options, A and B, we use a standard softmax function. The probability of choosing option A over B depends on the difference in utility between A and B in the following way:

$$p\left(chooseA\right) = \frac{1}{1 + e^{-\beta(U(A) - U(B))}}$$

where β is a noise parameter: the higher β, the less noise. If $\beta = 0$, the exponent term equals 1 and the agent chooses randomly between A and B. When the signed difference in utility between A and B is larger (more positive), the probability of choosing A is higher.

Each agent of each model responded 10 times to approximately 3000 trials (2934 to be exact). That is 30.000 responses per agent. Each model consisted of 1000+ agents ([1225,1225,1210] for each of the three models, to be exact). That is approximately 3000+ agents in total. With 30.000 responses each, we are talking about more than 90 million simulated data points. This required the use of the NYU's High Performance Computing (HPC) cluster.

2.4 Trial Selection 1.0: Independence and Uniform Priors

In our first attempt, we wanted to be able to treat each trial separately, so me assumed independence between trials. Strictly speaking, this is not correct if we are to select more than one trial for the experiment, which was our aim. That is because two trials may provide redundant information. The correct way of selecting n trials is by evaluating sets of n trials, but that would have incremented the computational demand exponentially. We resorted to assuming independence and evaluating model recovery performance, using simulated data, to validate our approach.

For the first and second iterations we also assumed uniform priors over models, and over agents within models. Each of the three models was presumed to be equally likely, at a probability of 1/3 each. And each of the agents was equally likely, at around 1/1000+. As we move along this process and get data, we incorporate the results as priors on the following iteration.

With these assumptions, the math to compute mutual information between data and models is the following:

$$MI\left(r, M\right) = \sum_{M}\sum_{r} p\left(r, M\right) \log \frac{p\left(r, M\right)}{p\left(r\right) p\left(M\right)}$$

$$= \sum_{M}\sum_{r} p\left(r, M\right) \left[\log \frac{p\left(r, M\right)}{p\left(r\right)} - \log p\left(M\right)\right]$$

$$= \sum_{M}\sum_{r} p\left(r, M\right) \log \frac{p\left(r, M\right)}{p\left(r\right)} - \sum_{M}\sum_{r} p\left(r, M\right) \log p\left(M\right)$$

With the assumption of uniform prior over models, $p(M) = 1/n_M$, where n_M is the number of models (for us, three). Then, the expression for mutual information simplifies to:

$$
\begin{aligned}
MI\left(r,M\right) &= \sum_{M}\sum_{r} p\left(r,M\right)\log\left(M\left|r\right.\right) - \left(\log\frac{1}{n_M}\right)\sum_{M}\sum_{r} p\left(r,M\right) \\
&= \sum_{M}\sum_{r} p\left(M\right) p\left(r\left|M\right.\right)\log\left(M\left|r\right.\right) - \log\left(\frac{1}{n_M}\right) \\
&= \frac{1}{n_M}\sum_{M}\sum_{r} p\left(r\left|M\right.\right)\log p\left(M\left|r\right.\right) + \log\left(n_M\right) \\
&= \frac{1}{n_M}\sum_{M}\left\langle \log p\left(M\left|r\right.\right)\right\rangle_{r|M} + k
\end{aligned}
$$

where $\langle\cdot\rangle_{r|M}$ denotes an expected value with respect to the distribution $p(r|M)$. This distribution represents the probabilities that either option within the candidate trial is chosen under the model.

2.4.1 Model Selection - Trial selection for Study 1

The first iteration aimed at identifying the best performing model, regardless of agent. The sum over models therefore includes a sum over agents. To numerically approximate these probabilities, we simulated responses from each model, assuming a linear grid of parameter combinations within each model, which we will call agents. For the Rawlsian model, the grid consisted of 35 steps for β (between 1 and 10) and 35 steps for ω (between 0 and 1). For the Cobb-Douglas model, the grid consisted of 35 steps for β (between 1 and 10) and 35 steps for γ (between 0 and 1). For the Charness-Rabin model, the grid consisted of 10 steps for β (between 1 and 10), 11 steps for ρ (between -1 and 1), and 11 steps for σ (between -1 and 1). To account for decision noise, each simulated agent responded 10 times to the candidate trials.

We ranked the 2934 candidate trials by MI (Figure 2.3) and selected the 100 most informative ones (Figure 2.4). Those trials on average provided 0.214 bits of mutual information, whereas the average of all 2934 trials was 0.078 bits. Thus, by selecting the 100 most informative trials we on average gain 2.75 times more information than by selecting 100 trials randomly.

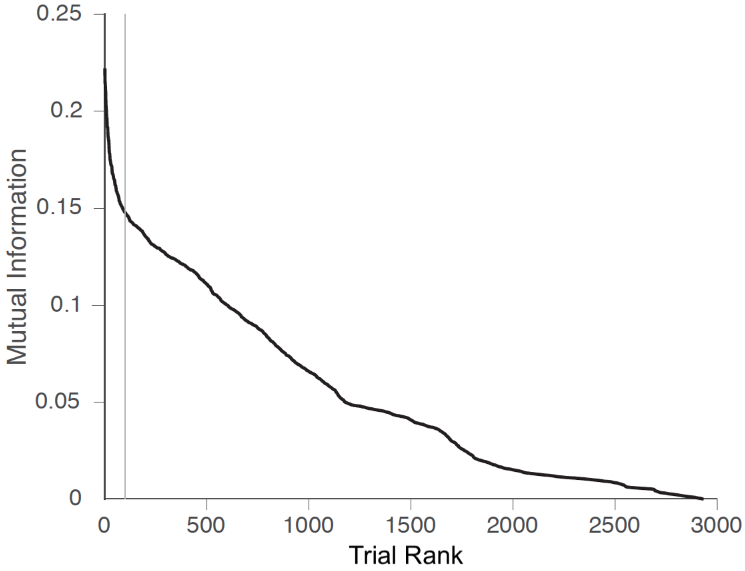

Figure 2.3 Mutual Information Decay. Decay in mutual information provided by most to least informative trials. Grey vertical line marks 100 trials

2.4.2 Model Recovery Performance

To validate our approach, we performed a model recovery test. For each model and each agent, we simulated responses to the 100 selected most informative trials. On each

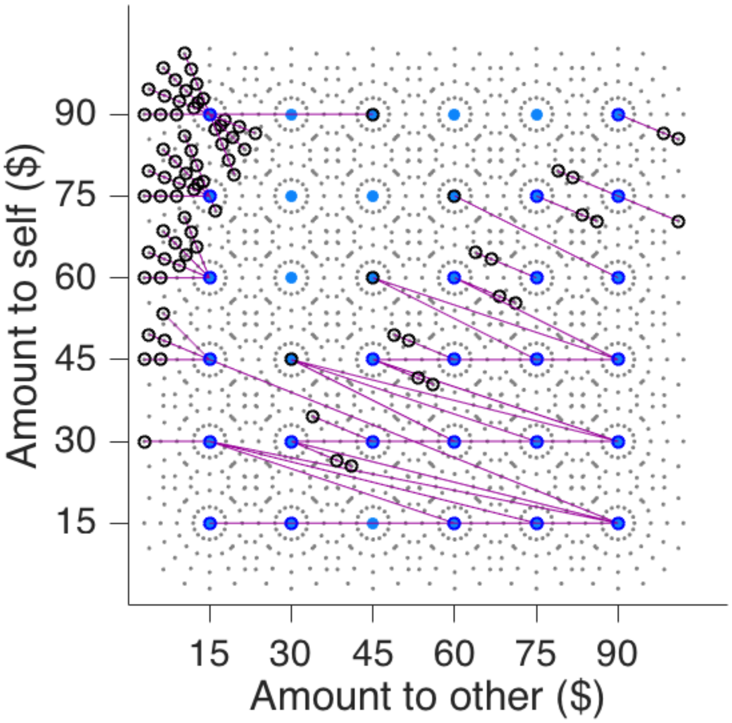

Figure 2.4 Selected Trials for Experiment 1. 100 most informative trials, selected to distinguish among candidate models. The two options of each trial are connected by a purple line

simulated data set generated in this way, we compared the three models using AIC and declared a winner. Model recovery performance is the percentage of data sets for which the winning model was equal to the true (generating) model. We repeated this procedure 10 times and computed the average for each model. Model recovery performance was above 90% for every model (see Table 2.1).

We repeated the above process with different numbers of most informative trials. Reducing the number of trials produced a marked drop in accuracy, while increasing the number of trials did not significantly increased accuracy. If we randomly chose 25 walrasian (positive-price) trials, model recovery performance drops to 45% for the Charness-Rabin model. Randomly choosing 25 walrasian trials (positive price) around the ($45, $45) reference point (the reference point that is closest to ($50, $50) does not help.

	Most informative trials				25 random walrasian trials	
	80	100	300	500	Any 25	Around 45,45
Charness-Rabin	0.70	0.91	0.92	0.92	0.45	0.38
Cobb-Douglas	0.96	0.96	0.96	0.96	0.89	0.85
Rawlsian	0.93	0.91	0.94	0.93	0.86	0.55

Table 2.1 Model Recovery Performance. Model recovery performance for simulated data sets of the three candidate models by number of selected 'most informative' trials (80, 100, 300, 500). A performance of 0.99 indicates almost perfect model recovery. Results reflect the average of 10 runs. For comparison, we include 25 random trials anywhere in the payment space and around the reference point $45,$45

2.4.3 Parameterization - Trial selection for Study 2

In Study 1, we selected trials to maximally distinguish between three models. In Study 2, we instead focused on estimating the two main parameters, ρ and σ, within the Charness-Rabin model (the best-fitting model in Study 1). We first examined how suitable the trials from Study 1 are for parameter recovery within this model. Recovery was good when either parameter was roughly in the range between (-0.2 and 0.6). Outside this range, parameter recovery faltered. Among the parameter combinations for which recovery was less good, there are some that are of particular conceptual interest:

- the classical difference aversion subspace (Fehr-Schmidt), where individuals are anti-social towards others when they are worse off but pro-social when they are better off ($\rho > 0$, $\sigma > 0$).
- the classical difference aversion subspace (Fehr-Schmidt), where individuals are anti-social towards others when they are worse off but pro-social when they are better off ($\rho > 0$, $\sigma < 0$).
- the competitive subspace, where individuals are anti-social towards others both when

they are worse off and when they are better off ($\rho < 0$, $\sigma < 0$);

To improve parameter recovery within these two subspaces, we defined within each subspace a new set of parameter combinations. In the classical difference-aversion subspace, ρ takes values from 0.1 to 1 in steps of 0.1, and σ takes values from -1 to -0.1 in steps of 0.1, for a total of 100 parameter combinations. In the competitive subspace, ρ takes values from -1 to 0 in steps of 0.1, and σ takes values from -1 to -0.2 in steps of 0.1, for a total of 99 parameter combinations. Then, again separately for each subspace, we computed the mutual information $MI(r, \theta)$ between the subject's response r and this new set. (This calculation is analogous to the trial selection process for Study 1 but with a parameter combination taking the place of a model.) We selected the top 25 most informative trials per subspace, for a total of 50 trials, which we added to the original trial set (Figure 2.5).

Given the combined set of 150 trials, parameter recovery was precise across the entire range of each parameter. Thus, we expect parameter estimates from Study 2 to be precise. See chapter 3 - A Systematic Characterization of Other-Regarding Preferences for the results.

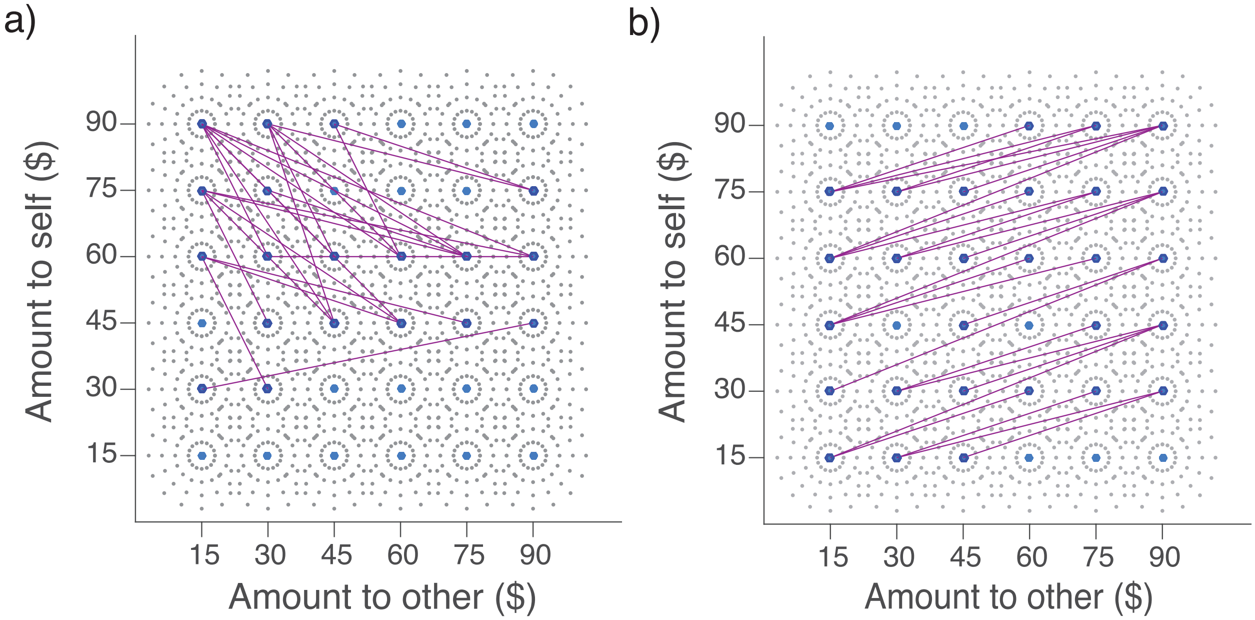

Figure 2.5 Additional Trials for Experiment 2. (a-b). Additional trials selected to distinguish subjects who are difference averse **(a)** and those who are competitive **(b)**. The two options of each trial are connected by a purple line

2.5 Trial Selection 2.0: Redundancy Reduction and Experimental Priors

At this stage we were able to make two significant improvements to the trial selection method. Firstly, we were able incorporate priors over types of agents. Not all agents are equally likely. From studies 1 and 2 we know which ones are. We know that there are a few competitive and difference-averse people, but most people are simply prosocial (see chapter 3 - A Systematic Characterization of Other-Regarding Preferences.

By tailoring the trials to better identify what we already suspected was there, we made the method even more efficient. This allowed us to reduce the number of trials needed to obtain the same amount of information and statistical power.

Secondly, we considered the potential dependency of the information provided by different trials. Independently ranked, the two most informative trials may provide very similar information. The correct way of choosing two trials is by evaluating sets of two, in every possible combination. For greater numbers, the computational demand increases exponentially.

We resorted to an iterative method, were on the first iteration we ranked each trial independently and selected the single most informative one. We used that as ‘history’ on the following iteration, where we again ranked each trial independently, but considering the response to the previously selected one. The winning trial on this iteration would then join the previous one as the latest addition to the ‘history’. On the third iteration, two trials will be already determined, and so on.

We evaluated the same 2934 candidate trials, but iteratively incorporating one trial at a time to an ever-increasing number of trials in the set. We simulated sets of 20, 50 and 75 trials. The first 20 trials of the 50 trials set size served as validation for the original 20 trials, and so on and so forth with the bigger sets. We settled on 75 trials, which we used in Study 3. For the results, see CH.E2 – The Effect of Cognitive Load on Social Preferences.

The math is as follows. Note that we approximated the distribution of agents by sampling, so the math for marginalizing over agents is equivalent to a uniform distribution.

$$MI\left(r_{history}, M\right) = \frac{1}{n_M} \sum_{M} \sum_{r_{history}} p\left(r_{history} \left| M\right.\right) \log p\left(M \left| r_{history}\right.\right) + k$$

$$\propto \frac{1}{n_M} \sum_{M} \sum_{r_{history}} p\left(r_{history} \left| M\right.\right) \log \frac{p\left(r_{history} | M\right) p(M)}{\sum\limits_{M} p\left(r_{history} | M\right) p(M)}$$

$$= \frac{1}{n_M} \sum_{M} \sum_{r_{history}} \prod_{t=1:\tau} p\left(r_t \left| M\right.\right) \log \frac{\prod\limits_{t=1:\tau} p(r_t | M)}{\sum\limits_{M} \prod\limits_{t=1:\tau} p(r_t | M)}$$

$$= \frac{1}{n_M} \sum_{M} \sum_{r_{history}} p\left(r_\tau \left| M\right.\right) \prod_{t=1:\tau-1} p\left(r_t \left| M\right.\right) \log \frac{p(r_\tau | M) \prod\limits_{t=1:\tau-1} p(r_t | M)}{\sum\limits_{M} p(r_\tau | M) \prod\limits_{t=1:\tau-1} p(r_t | M)}$$

2.5.1 Experimental Priors and Simulated Agents

We used the data obtained in Study 2, which is precise in terms of the parameter estimates within the Charness-Rabin model, to create the simulated agents that this iteration of the DGM method will aim to distinguish. The simulated agents reproduced the experimental distribution by sampling.

We had date from 30 subjects who participated in Study 2. Each subject provided 2 pairs of rho and sigma parameters, one for friend and one for stranger. In total, we had 60 observations. We separated each observation into one of three quadrants: prosocial (positive rho, positive sigma), difference-averse (positive rho, negative sigma), and competitive (negative rho, negative sigma). 75% of observations were in the prosocial quadrant, 15% in the difference-averse one, and 10% in the competitive quadrant. We worked with each quadrant separately because there is a certain qualitative difference in behavior represented by each one. Small quantitative differences between agents across quadrants are, to us, more significant than greater quantitative differences between agents within a given quadrant.

Because agents within each quadrant were bounded by the limits of the quadrant, we could not fit a Gaussian function to approximate their distribution. Instead, we fitted a two-dimensional lognormal distribution by taking the absolute values, then taking logarithm, and finally taking their mean and covariance. Using those statistics, we generated 300 random agents from each distribution corresponding to each quadrant. We then randomly selected the corresponding proportion (75%, 15%, and 10%) of agents so that there would be a total of 300 simulated agents (instead of 900) that represented the experimental distribution. Finally, each agent was converted back to the original range, by taking the exponent and applying the appropriate signs, according to each quadrant.

2.5.2 Trial Dependence - The Iterative Solution

We considered the same 2934 candidate trials used when we first started developing this method. In this instantiation, we evaluated sets of 20, 50 and 75 trials, on each case iteratively incorporating one trial at a time to an ever-increasing number of trials in the set, until the desired number of selected trials was reached.

On the first iteration, every trial was evaluated independently and the most informative one was selected to integrate the history of responses, then removed from the pool of trials to be evaluated next. On the following iterations, the remaining trials were evaluated independently of each other, but considering the response to the trials in the history.

While each of the remaining trials were evaluated independently, by considering the history of responses to previously selected trials, we were indeed taking into account the dependence of responses to trials already in the set. In effect, on the nth iteration, we were calculating the information provided by sets of n trials, n-1 of which were the same across all sets, and only the nth one differed.

It should be noted that, on each iteration, an entirely new set of simulated data was created to guarantee that the results and selected trials would be independent of any particular simulation.

2.5.3 Performance Analysis

To evaluate the performance of our method and determine how many trials to use in our next experiment, we constructed realistic data for each trial set size. Each of the 300 simulated agents responded one time to each of the selected trials, first 20, then 50, then 75. These realistic data sets were fitted as real data would be, using the bads algorithm developed by Luigi Acerbi at the Ma lab (Acerbi and Ma, 2017). This process was repeated 10 times, and the fitted parameters averaged for each agent. We then compared the fitted parameters to the true, generating parameters, and calculated the error as the distance between the two.

As we can see in Figure 2.6, performance improves significantly from 20 to 50 to 75 trials. In black circles we can observe the true, generating parameters of our simulations. In grey circles, we see the fitted parameters. For each agent, there is a line connecting the true and the fitted parameters. The color hue codes for the error between the two. In Figure 2.6(a-c), we see the results for 20, 50, and 75 trials respectively, using the 2nd generation trials on simulated agents distributed following the experimental results from previous studies. Figure 2.6(d-g) show the results for 20, 50, 75, and 150 trials respectively, using the 1st generation trials on the same experimental agents, despite the fact that those trials were selected aiming to distinguish a uniform distribution of simulated agents. But even if we evaluate the 1st generation trials on the uniform agents they were designed to distinguish (Figure 2.6(i)), the 2nd generation trials vastly outperform them. In Figure 2.6(h) we see the performance of the trials used by Hutcherson et al, 2015, which disputes that the status-dependent Charness-Rabin model outperforms a simple linear model.

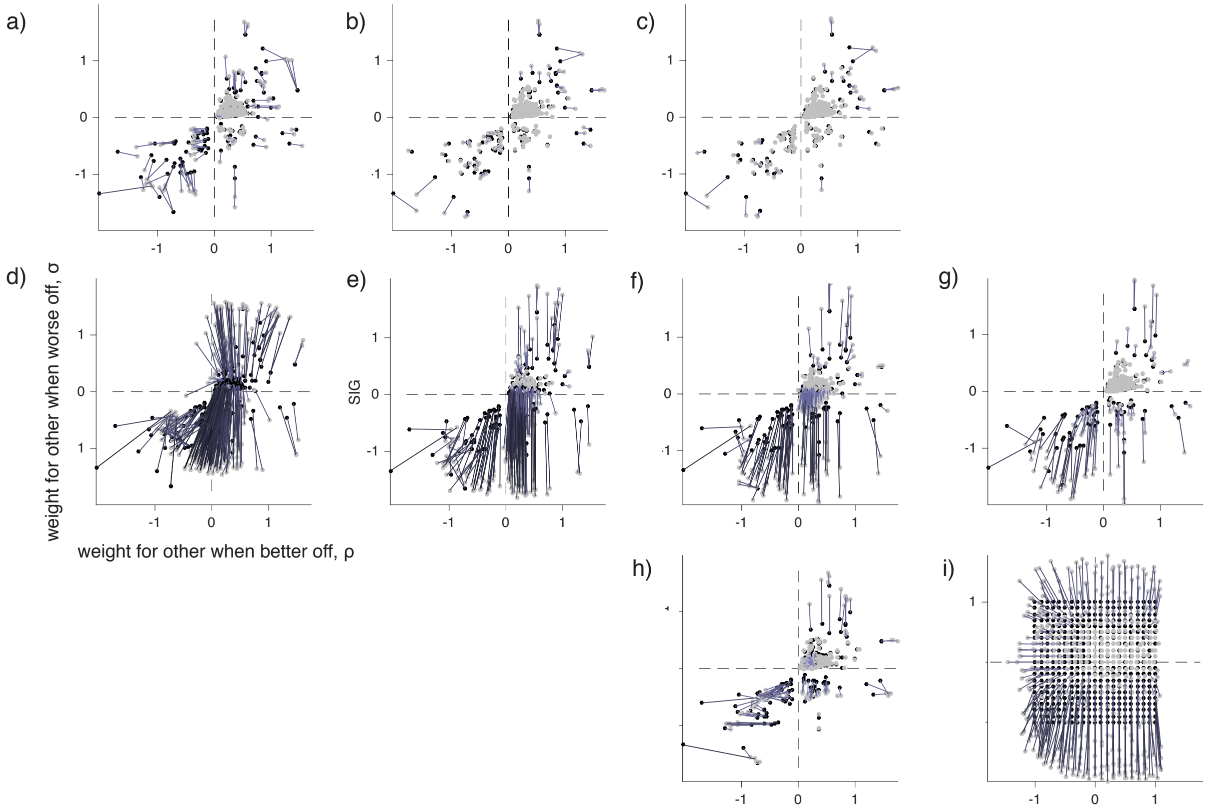

Figure 2.6 Visualization of Trial Performance. In every panel, true generating parameters are depicted in black circles, and estimated parameters in grey. A color coded line connects the true and estimated parameters that correspond to the same agent. The darker the line, the greater the distance between the two **(a-c)** Performance of 20, 50 and 75 2nd generation trials.**(d-g)** Performance of 20, 50, 75 and 150 1st generation trials. These trials were obtained targeting uniformly distributed simulated agents but here we present their performance with agents that follow the experimental distribution. **(h)** Performance of 80 trials used by Hutcherson et al, 2015. **(i)** Performance of 150 1st generation trials on uniformly distributed agents.

In terms of the root mean square error (RMS), we can see in Figure 2.7 that the error between fitted and true parameters systematically diminishes as we increase the number of trials both within the 1st and 2nd generation trials, but the 2nd generation ones consistently produce smaller errors than the 1st generations ones.

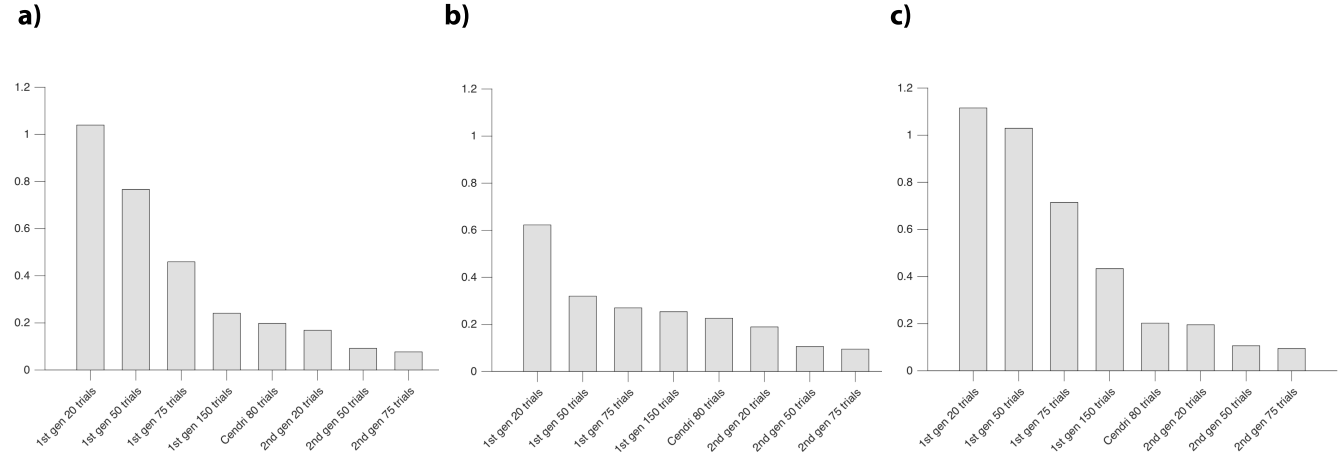

Figure 2.7 Quantification of Trial Performance. **(a)** Root Mean Square (RMS) errors produced by different sets of trials considering both parameters **(b)** the ρ parameter only, and **(c)** the σ parameter only. In every case, the trial sets under evaluation are the 1st generation trials (evaluated on experimentally distributed agents), the 2nd generation trials, and I include for comparison the trials used by Hutcherson et al, 2015, which are referred to as Cendri trials.

We proceeded to use 75 2nd generation trials for our following experiment, which consisted of a cognitive load task. See chapter 4 - The Effect of Cognitive Load on Other-Regarding Preferences. The selected trials are presented in Figure 2.8.

2.6 Conclusion

In this chapter we have presented the development of a novel trial selection method that aims to overcome the ad hoc nature of trial selection used in experiments so far.

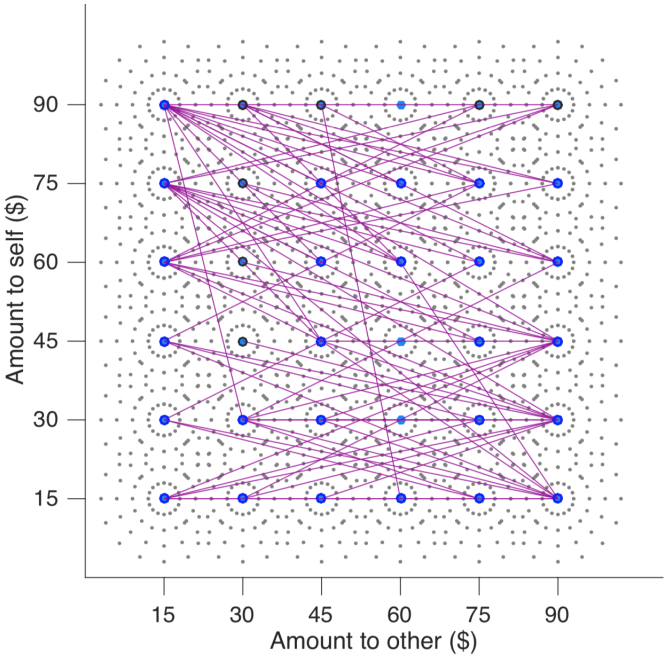

Figure 2.8 2nd Generation Trials. 75 selected trials

.

There has been enormous progress in testing since the original dictator game, and we can see researchers intuitively progressing towards trials that extract more information about the hidden variables of interest. Here we propose a systematic approach to do just that.

This method is based on information theory. We used simulations to compute the mutual information between unobservable preferences and observable choices to a vast number of candidate trials. Trials high in mutual information between these two variables will produce choices that reduce the most amount of uncertainty about the preferences that have produced them.

We developed this method in stages, first assuming independence of responses to different trials and uniform priors over models of preferences. This stage produced the trials selected for study 1, aimed at identifying the best performing model, and those selected for study 2, which parametrized the winning model from study 1. The results of these two studies are presented in chapter 3 - A Systematic Characterization of Other-Regarding Preferences.

On the second stage, we incorporated dependence of responses to different trials and experimental priors over preferences obtained after the first stage. Incorporating dependence was especially challenging due to the prohibitive computational demand of considering sets of trials, in every possible combination, and of differing sizes. We were able to overcome this challenge by applying an iterative approach, starting with a set of 1 trial, then 2, then 3, and so on. This approach proved effective and allowed us to extract more information with as few as half the trials compared to the previous stage. This stage produced the trials used in study 3. The results of this study are presented in chapter 4 - The Effect of Cognitive Load on Other-Regarding Preferences.

The specific trials selected for our studies during this process depend on i- the original set of candidate trials and ii- the preferences we were trying to identify (candidate models). But the method is general and can be applied to different problems, or even the same problem better formulated. Another pool of candidate trials might throw better results. Another list of candidate trials may find a better description of our preferences for generosity. We encourage others to continue the task, both by improving the trial selection method, and by testing other models and other trials.

Chapter 3

A Systematic Characterization of Other-Regarding Preferences

3.1 Introduction

With the description of the trial selection method on the previous chapter behind us, I turn to the experimental questions that motivated this thesis in the first place.

Humans, like many other animals, are a social species. As such, they develop emotional connections to others, mostly positive, but sometimes negative as well. Positive emotional connections lead us enjoy our friends' happiness and worry about their sorrows while negative emotional connections might lead us to resent our foe's victories and savor their defeats (Singer et al., 2006; Smith, 1759). These connections to others can serve as a motivation to sacrifice personal benefit for the benefit of our friends and perhaps as well to incur costs to hinder our foes.

In a groundbreaking experiment, Andreoni and Miller showed that giving behavior towards strangers may be rational, and that for the most part, it is. Thus, Homoeconomicus, the rational man, need not be selfish (Glimcher et al., 2005; Henrich et al., 2001). The implication of altruism as rational behavior is that we can understand it as arising from the maximization of a utility function, a process by which a decision function ranks options according to our preferences and selects the best. The specific form such a function takes is a postulate about our inclinations to consider others in our decisions and constitutes a model of other-regarding preferences. There are multiple such models (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Cox and Sadiraj, 2005; Fehr and Schmidt, 1999; Levine, 1998), but there is no consensus about which one describes attitudes towards others best. Our goal here is to characterize other-regarding preferences towards friends and strangers by formally comparing several candidate models on an individual-subject basis and by examining the parameters of the best-fitting model.

3.2 Methods

3.2.1 Task: the Dictator Game

We used a two-alternative forced-choice dictator game in which the recipient was either a friend of the subject or a stranger. Each option consisted of an amount of money to Self and an amount of money to Other (Figure 3.1). Previous studies have used variants of dictator games to elicit other-regarding preferences (Andreoni and Miller, 2002; Charness and Rabin, 2002; Cox and Sadiraj, 2005; Engelmann and Strobel, 2004; Fisman et al., 2007; Forsythe et al., 1994). In such games, the recipient does not have any role in determining the outcome, and therefore the scope of strategic thinking by the decision-maker is minimized. Thus, related but separate processes, such as reciprocity, can arguably be ruled out.

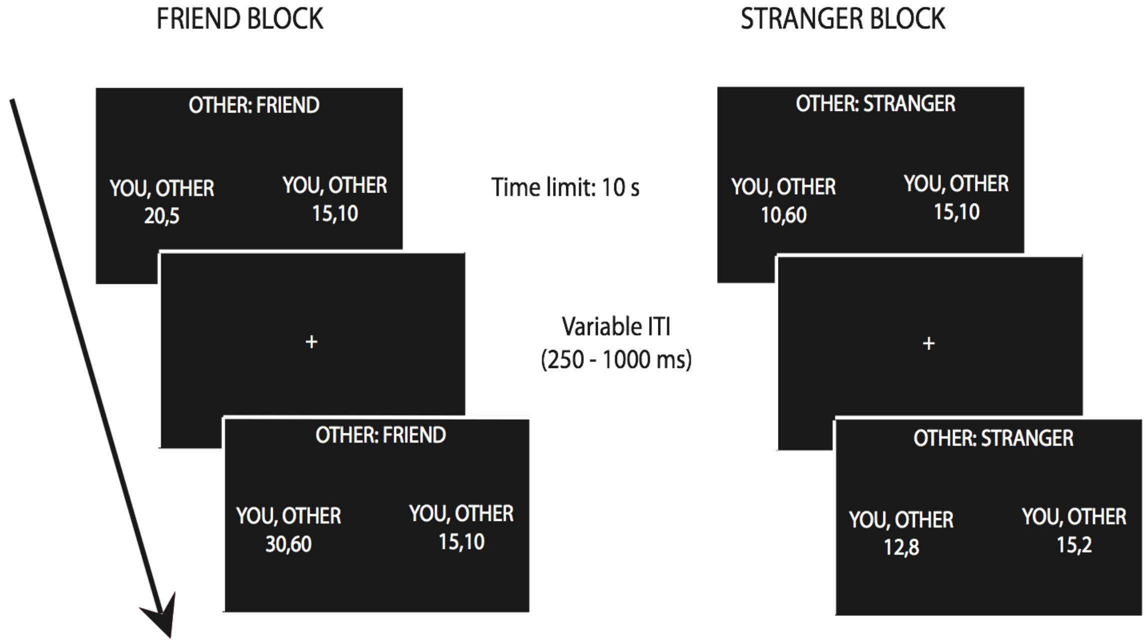

Figure 3.1 Dictator Game Task. Each trial consists of two options. Each option contains two amounts, one for self and one for other. The other can be a friend of the decision maker or a stranger, according to block.

3.2.2 Incentive Compatibility and Collusion Avoidance

We implemented several procedures to minimize the effect of strategic thinking, reputation concerns, and social coercion for the decision-makers, and to discourage collusion between decision makers and their friends. At the end of the experiment, one “friend” trial and one “stranger” trial were randomly selected, and payment was implemented according to the decision-maker’s choices on those trials. This procedure was meant to elicit incentive-compatible behavior. For the friend and the stranger, the decision-maker’s choice on the corresponding trial was realized for payment. For the decision-maker themselves, only one of the two trials randomly selected was realized for payment in order to maintain everyone on the same range of potential earnings. Nobody (decision-maker, friend, or stranger) was informed of the non-chosen option on the trials selected for payment. This procedure made it impossible for the friend to know whether the decision-maker had been generous or selfish (the non-chosen option could have been much worse for the friend and much better for the decision-maker, or not). Payments were made to each participant without the presence of others, and nobody was informed of any of the other participants' payoffs. We asked each participant to mix the payment bills with the bills they already had in their wallets before leaving the payment room. Thus, any sharing of payment information or collusion had to rely on trust.

Before starting the experiment, we explained to all participants the aforementioned procedures and why they were put in place. Specifically, we explained i- that only one "friend" trial and one "stranger" trial would be selected for payment and that the decision-maker's choice in that trial would be realized, so the decision-maker should treat each trial as if it were the only one, and ii- that none would be informed of the non-chosen payoff on the selected trials, so friends could not read into the amounts they received, and therefore could not get offended for receiving too little. The decision maker should therefore not worry about what their friend might think.

3.2.3 Rationality Tests

Rationality is not an all-or-nothing variable. Violations of rationality can be more or less severe, meaning that people may depart from perfect rationality in a greater or lesser degree. We use Afriat's Critical Cost Efficiency Index (CCEI) as a measure of the extent to which subjects depart from perfect consistency (Afriat, 1972). Afriat's index calculates how much the choice/budget sets must be adjusted to remove all GARP violations. GARP stands for Generalized Axiom of Revealed Preference and it constitutes the behavioral counterpart of the mathematical representation of rationality (Afriat, 1967) (See also the Introduction, subsection: The Economics in Neuroeconomics, for a more detailed explanation of rationality). The index goes from 0 to 1, with numbers closer to 1 indicating greater adherence to GARP, or a greater level of rationality. We analyzed rationality only in the walrasian subspace of trials (trials with a positive price, or negative slope). In the experiment, we also provided non-walrasian trials, for which the rationality index cannot be calculated. We assume the same level of consistency extends to the non-walrasian subspace.

3.2.4 Collusion Tests

If they are indeed colluding, we should see that the decision-maker chooses the option with the greater total sum every time, or if we allow for imperfect behavior, almost all of the time. In addition, if the decision maker and the friend were colluding, we would expect to find purely utilitarian preferences, i.e. money to themselves is equally valued to money to their friend. This would be manifested in equal parameter estimates for Self and Other.

3.2.5 Models of Other-Regarding Preferences

We considered three models that represent three different decision-making mechanisms: a generalized Rawlsian model, a Cobb-Douglas model, and the Charness-Rabin model (Charness and Rabin, 2002; Rawls, 1971). The classic difference-aversion model, or Fehr-Schmidt model (Fehr and Schmidt, 1999), is a special case of the Charness-Rabin model, and therefore, we only considered the latter for analysis. For their mathematical formulation, see chapter 2 - An Information Efficient Trial Selection Method. In the following sections, we will describe what they represent in terms of preferences.

3.2.5.1 The Rawlsian Model

The Rawlsian model was inspired on the philosophical work of John Rawls (Rawls, 1971). Rawls started by asking, if we were all waiting to be born into this world and didn't know in which circumstance that will be, what wealth distribution would we vote for? He speculated that we would want absolute equality, which would be achieved by 'maximin' preferences. A maximin preference maximizes the welfare of the worst-off individual in society. It is unlikely that our preferences would shift from considering solely one individual to considering solely another one, but because these preferences represent a radical egalitarian, we considered it interesting to include this model. We modified it slightly to allow for a degree of 'self-pity', or egocentric predilection, while maintaining the radical shift mechanism. A pure rawlsian agent is still discoverable in this modified version.

3.2.5.2 The Cobb-Douglas Model

The Cobb-Douglas model is a standard utility function used in introductory economics classes. It captures several key features widely believed to intuitively represent preferences. One such feature is diminishing marginal returns. This means that the more I have of a certain good, the less I value it relative to another one. Thus, I would be willing to give up more of it to get one unit of the scarce good. The effect of this feature is quite moderate preferences. Extremes where there is only one type of good and nothing of the other good would never be chosen under Cobb-Douglas preferences. The parameters of the Cobb-Douglas preference represent how much each good is valued, but they are restricted to be positive numbers, so when it comes to other-regarding preferences, this feature can be quite limiting.

3.2.5.3 The Charness-Rabin Model

The Charness-Rabin model overcomes this limitation. It posits a linear combination of self and other with the weight on the other depending on who is better off. The linearity of this preference is less realistic than the convexity of the Cobb-Douglas, but this is compensated by the fact that the parameter indicating how we value others can now be negative, introducing an entirely new range of possible emotions towards others. It also brings up an interesting feature for other-regarding preferences: relative status.

The Charness-Rabin model shares the relative status feature with the Fehr-Schmidt model, which arises from modelling difference-averse preferences. These are preferences that increase with one's own wealth, but decrease, differentially, with inequality. They decrease with advantageous as well as disadvantageous inequality, with the latter being more bothersome than the former. In the end, the mathematical form of the two models is very similar, a linear combination of self and other, with the weight on the other depending on who is better off. But, by virtue of positing difference-aversion, the Fehr-Schmidt model imposes an antisocial disposition towards the other when the decision maker is worse off. The Charness-Rabin model allows for this antisocial disposition, but it does not impose it.

In this sense, the Fehr-Schmidt model is a special case of the Charness-Rabin model, and therefore, we only include the latter in the list of candidate models. But, as in the case of the Rawlsian preference, the Fehr-Schmidt preference is still relevant in what it represents, and we will be looking at the parameters to see this subspace is where most Charness-Rabin agents lie.

3.2.6 Mapping Utility to Choice

In all three models, we included softmax noise to account for variability in human decisions. As mentioned in Chapter 2 - An Information Efficient Trial Selection Method, the probability of choosing option A over B depends on the difference in utility between A and B in the following way:

$$p\left(chooseA\right) = \frac{1}{1 + e^{-\beta(U(A) - U(B))}}$$

where β is a noise parameter: the higher β, the less noise. If $\beta = 0$, the exponent term equals 1 and the agent chooses randomly between A and B. When the signed difference in utility between A and B is larger (more positive), the probability of choosing A is higher.

3.2.7 Trial Selection

Prior to running the experiment, we developed a novel trial selection method aimed at maximizing the amount of information that could be extracted in a single experimental session. We describe this method in detail in Chapter 2 - An Information Efficient Trial Selection Method.

To summarize here, this method uses simulated data to maximize the mutual information between the models of interest and the responses these models produce to different candidate trials. Mutual information is a measure of mutual dependence, meaning that knowing about one variable reveals something about the other variable (Cover and Thomas, 2012). In this case, we were looking for trials the responses to which would reveal the most amount of information about the underlying preferences (models) that had produced those choices. We developed this method iteratively, incorporating the data obtained in earlier stages as priors on the following ones. This provided improved trials as we progressed on the experimental projects. We first aimed to identify the best performing model, regardless of agent. This produced 100 selected trials for study 1. Once the best performing model was identified (see section Results of this chapter), we added 50 trials aimed at improving parametrization of the winning model, specifically in two subspaces of the parameters' domain: the difference-averse and the competitive agents. This produced 150 trials used in study 2. See Figures 2.4 and 2.5(a-b).

The method continued improving, providing ever better trials, which we used in future projects 4 - The Effect of Cognitive Load on Other-Regarding Preferences.

3.2.8 Model Fitting and Model Comparison

We fitted the parameters of each model for each individual subject using maximum-likelihood estimation. We wrote custom code for the likelihood function over the parameters and maximized the likelihood by using a genetic algorithm (Matlab's ga function). For model comparison, we used two common metrics for goodness of fit, both of which penalize for model complexity: the Akaike Information Criterion (AIC) and the log Bayes factor (Akaike, 1974; Kass and Raftery, 1995). Denoting the data by D and the ith model by Mi, the log Bayes factor of the i^{th} model relative to the j^{th} model is:

$$LBF_{i,j} = \log \frac{p\left(D\left|M_i\right.\right)}{p\left(D\left|M_j\right.\right)}$$

Here, the marginal likelihood of model M is an integral over the model parameters θ:

$$p\left(D\left|M\right.\right) = \int p\left(D\left|\theta, M\right.\right) p\left(\theta\left|M\right.\right) d\theta$$

We approximated this integral as a sum over a dense grid of parameter combinations (1 million in total). For each parameter, we assumed a discrete uniform prior over the corresponding grid dimension. When comparing models, we assumed that each subject follows the same model, regardless of partner.

3.2.9 Group Analysis

To evaluate whether there is a model that best describes the entire population or a distribution of models that describe different types of individuals, we turn to a hierarchical Bayesian approach that allows us to perform inference in model space (Stephan et al., 2009). Instead of assuming that one model is true for the entire population and selecting the one that performs better under this assumption, we assume there is a multinomial distribution of models. The probability that any random subject follows one particular model would be the frequency of that model in the population. We then use the log evidence in favor of each model, for all subjects, to estimate the multinomial probabilities of the models in the population.

3.3 Results

3.3.1 Experiment 1: Model Selection

3.3.1.1 Subjects

Thirty subjects participated in Experiment 1 (7 females, 23 males; median age 22.5). The study conformed to the Declaration of Helsinki and was approved by the Institutional Review Board of New York University. The subjects completed the same 100 trials twice, once with the friend and once with the stranger, in alternating blocks of 25 trials in which the other – friend or stranger- remained the same. Both trial order and block order were randomized. We provided a neutral context and implemented several procedures to discourage self-imposed social coercion, reputation concerns, and collusion between decision-makers and their friends (see this chapter's Methods section, subsection Incentive Compatibility and Collusion Avoidance). Subjects, friends, and strangers remained in different rooms of the lab for the duration of the experiment to avoid decisions based on different downtimes.

3.3.1.2 Rationality and Collusion

Afriat's index of rationality is high across the board, and it maintains high for the duration of the experiment. The minimum average index score was observed in block 4 at a value above 0.85, or 85% adherence with GARP. See supplementary figure S1.A. This validates our utility function approach. In terms of collusion with friends, the maximum proportion of trials in which a subject chose the option with the maximum total sum was

64 out of 96 trials (for this subject, we discarded 3 trials because of the display problem and one trial that timed-out before they could make their choice). Even this subject is far from serious collusion measured as systematically choosing the option with the greater total sum, so we discarded the concern that our participants might be trying to game us into maximum payout. This was confirmed once we performed the model fits and compared our winning model to a pure utilitarian agent. The pure utilitarian agent lost to our winning model in 26 out of 30 cases. See supplementary info. Together, these two analyses suggest that the mechanisms we put in place to deter subjects from colluding with friends worked.

3.3.1.3 Model fits and psychometric curves

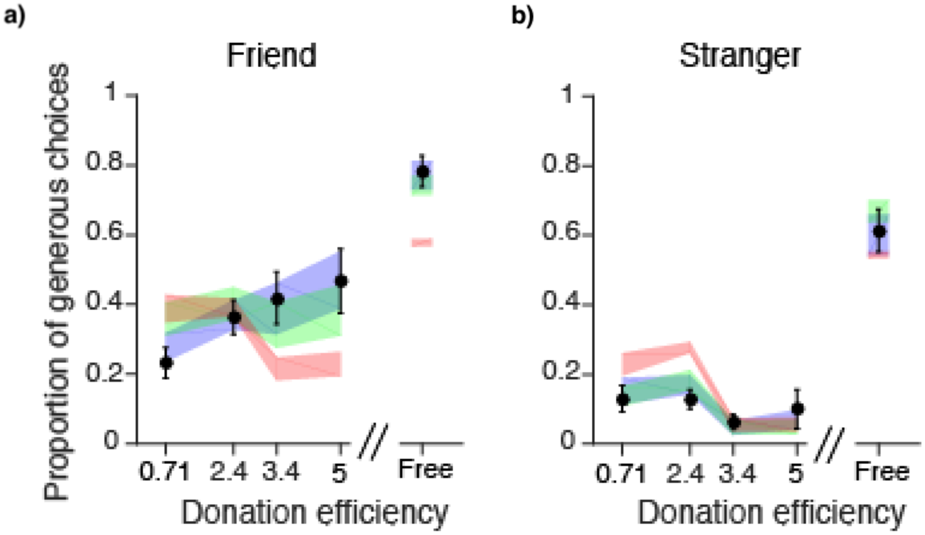

Figure 3.2 Psychometric Curves (a-b). Psychometric curves for decisions involving **(a)** friends and **(b)** strangers, depict the proportion of generous choices at 4 levels of donation efficiency and a fifth level where donating is free (donation efficiency is ∞). Black dots and error bars represent the mean $\pm$ s.e.m. of the data; shaded areas represent mean $\pm$ s.e.m. of the fitted models. Blue=Charness-Rabin; Green=Cobb-Douglas;Red=Rawlsian models

We defined an option as generous if in that option, the other was better off than in the alternative option. In most trials, the subject was better off in one option and worse off in the other option, so choosing the generous option entailed a measure of sacrifice. We defined donation efficiency as the amount of money the other gets for each dollar the subject forgoes; the higher the donation efficiency, the cheaper it is to make the other receive an extra dollar. We examined the proportion of generous choices as a function of donation efficiency (Figure 3.2). When it comes to friends, the psychometric curve shows a steady increase in generosity as generousity becomes cheaper. This continues at extremely high donation efficiencies, in spite of the fact that the friend may end up ahead of the subject in terms of final payment. When it comes to strangers, generosity was low across the board except when being generous was 'free', i.e. when the amount to oneself was equal in both options, and only the other's amount differed. A two-way repeated-measures ANOVA on the proportion of generous choices revealed a significant main effect of donation efficiency (F(4,116)=36.03, p<10-14), a significant main effect of partner (F(1,116)=25.86, p<10-4), and a significant interaction (F(4,116)=5.13, p<10-3).

We fitted the three candidate models – Rawlsian, Cobb-Douglas, and Charness-Rabin – to individual-subject data (separately for friend and stranger) using maximum-likelihood estimation. Examining the fits of the three models to the psychometric curves, we observe that for both friends and strangers, the Charness-Rabin model follows the data more closely than the Cobb-Douglas and the Rawlsian models. Interestingly, the fitted Charness-Rabin model correctly predicts that generosity towards friends increases steadily with donation efficiency. By contrast, the two other models incorrectly predict a dip in the proportion of generous choices as donation efficiency increases and the decision-makers start to lag behind the other in earnings, a pattern more consistent with classic difference-aversion.

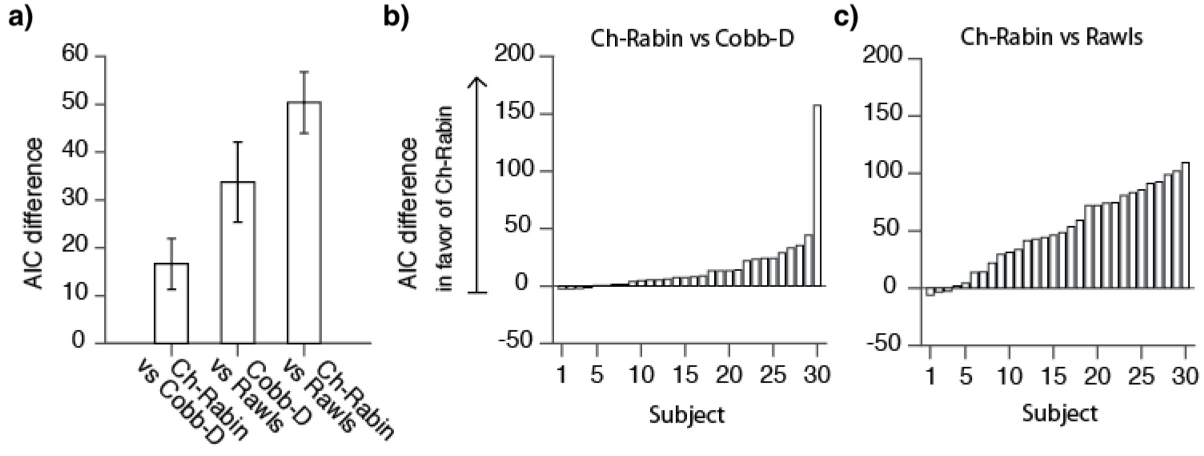

Figure 3.3 AIC Model Comparisons. **(a)** Mean AIC differences for pairwise model comparisons with bootstrapped error bars. Positive values mean that the model mentioned first in the label fitted better. **(b-c)** Individual AIC differences ranked from subjects showing least strength of evidence to most in favor of Charness-Rabin compared to **(b)** Cobb-Douglas and **(c)** Rawlsian models

For a quantitative measure of goodness of fit, we compared the models based on the AIC (Figure 3.3). We assumed that subjects' decision-making mechanism -the model- would be the same for friends and strangers. This assumption allowed us to add the AIC of each model across friend and stranger data. For 26 out of 30 subjects, Charness-Rabin had a lower (better) AIC than both the Cobb-Douglas and Rawlsian models. For the 4 subjects for whom the Cobb-Douglas or the Rawlsian models had a lower AIC than the Charness-Rabin model, the difference was not large (<6.0). The AIC of the Charness-Rabin model (summed across subjects) was lower than that of the Rawlsian model by 1511 (bootstrapped 95% confidence interval: [1135 1892]), and lower than that of the Cobb-Douglas model by 499 (*CI*: [260 870]). The AIC of the Cobb-Douglas model was lower than that of the Rawlsian model by 1012 (*CI*: [480 1468]). Here and elsewhere, we also estimated log marginal likelihoods (log Bayes factor) as an alternative to AIC. The results were consistent (see Supplementary Information).

Because the Charness-Rabin model was the only linear model under consideration, with its interesting feature being that it postulates a different weight on the other depending on whether the decision-maker is better or worse off, we asked whether a simple linear model, in which the weights to Self and Other are the same regardless of relative status, would do as well as the Charness-Rabin model. We found that this was not the case: AIC of the Charness-Rabin was lower than that of the simple linear model for 18 out of 30 subjects by 316 (*CI*: [48 748]).

3.3.1.4 Group Analysis

The AIC analysis implicitly assumes that all subjects follow the same model. However, we could instead allow for the possibility that all three models were represented in the population. We used a hierarchical Bayesian method to estimate the prevalence of each model in the population (Stephan et al., 2009). We found that the Charness-Rabin model best describes 94% of the population, the Cobb-Douglas model 3%, and the Rawlsian model 3%. The probability that the Charness-Rabin model is the most prevalent model in the population is greater than 0.99.

3.3.2 Experiment 2: Parameterization

3.3.2.1 Subjects

Whereas Experiment 1 taught us about the best model to describe human preferences for generosity, it did not emphasize individual differences within models. Experiment 2 was aimed at improving parameter estimation within the Charness-Rabin model, as opposed to model selection. Thirty subjects completed the task with these 150 trials twice, once with the other being a friend and once with the other being a stranger (13 females, 17 males; median age 23).

3.3.2.2 Rationality and Collusion

Rationality was high across the experiment, although we observed lower dips and more variability in this experiment compared to experiment 1. This might be explained by the fact that this session was 50% longer. Some participants may have experienced a degree of fatigue. The minimum index score was observed in the very last block, at a value of 0.78. This is still high, at almost 80% adherence to GARP.

In terms of collusion, one subject chose the option with the greatest total payment 70% of the time, followed by another subject at 60%. This may very well be explained by their preferences. If they were actively colluding, we would expect a higher proportion of choices of the greatest total payment. We discarded this concern.

3.3.2.3 Model Fits

Model comparison results were consistent with Experiment 1 (see Supplementary Material). A 2-way repeated measures ANOVA on the parameters showed a significant main effect of relative status (better off or worse off) ($F(1,29)=20.64$, $p=0.0001$), a significant main effect of partner ($F(1,29)=16.21$, $p=0.0004$), and no significant interaction ($F(1,29)=3.45$, $p=0.0734$).

The mean value of ρ (weight for other when better off) was positive and significantly

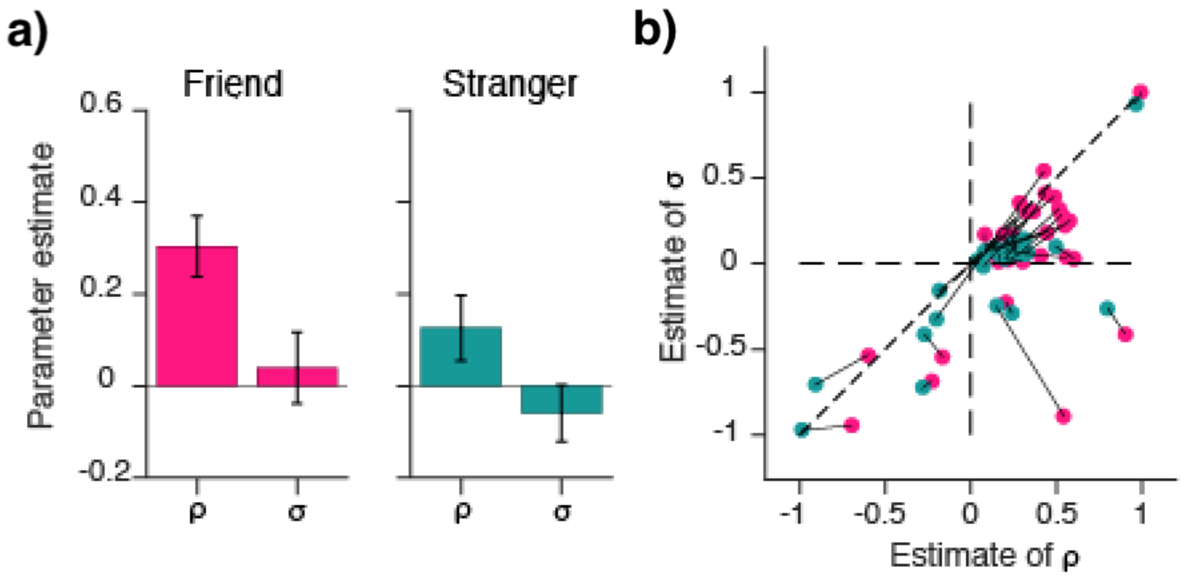

Figure 3.4 Other-Regarding Preferences, at the Individual and Group Levels. **(a)** Mean parameter estimates towards friends (red) and strangers (green) when better off (ρ) and worse off(σ), ± s.e.m. **(b)** Individual parameter estimates towards friends (red) and strangers (green). Parameters towards friend and strangers corresponding to the same subject are connected by a grey line.

different from zero when the other was a friend (M=0.304, SD=0.363, SEM=0.066; t(29)= 4.584, p<10-4) and positive but only marginally significant when the other was a stranger (M=0.126, SD=0.391, SEM=0.071; t(29)= 1.769, p=0.087). The mean value of σ (weight for other when worse off) was positive but not significantly different from zero when the other was a friend (M=0.039, SD=0.426, SEM=0.078; t(29)=0.504, p=0.62) and negative but not significantly different from zero when the other was a stranger (M=-0.059, SD=0.341, SEM= 0.062; t(29)= -0.954, p= 0.35).

These results indicate that on average, subjects tend to display prosocial preferences towards friends, with positive weights for Other both when better off and worse off, albeit less so when worse off. Prosociality may drop markedly when worse off, and even get close to indifference (weight of 0 for the other), but on average it does not go all the way to the realm of negative numbers, which would reflect competitiveness or antisocial preferences. When it comes to strangers, subjects on average tend to display difference-aversion preferences, with positive weights for the other when better off and negative weights for the other when worse off, albeit low in magnitude. This means that, when it comes to strangers, on the aggregate, there is a switch from prosocial to antisocial behavior when the relative status of the decision maker is disadvantageous (Figure 3.4(c)).

However, looking at the parameters at the individual level revealed a completely different picture. Not only were most subjects strictly prosocial; we found that other-regarding preferences towards friends and strangers tended to vary in degree more than in kind. At the individual level, those who were prosocial towards friends were also prosocial towards strangers, although less so. Those who were competitive towards strangers were typically competitive towards friends as well, and those who were difference-averse, i.e. prosocial when better off but antisocial when worse off- were consistently so towards friends as well as strangers. Of the 30 subjects on Experiment 2, 21 were 'prosocial' ($\rho > 0$ and $\sigma > 0$) both towards friends and strangers, and both when better off and worse off. When worse off, the weight for Other could diminish markedly, particularly in the case of strangers, but it typically did not go all the way to the negative realm (something akin to a 'do no harm' rule). Only two subjects were prosocial towards friends and switched to competitive towards strangers. Of our sample of 30 subjects, only three were classically difference-averse ($\rho > 0$ and $\sigma < 0$) and they displayed this pattern both towards their friends and the strangers, meaning that regardless of partner, their decision-making pattern was a combination of prosocial and antisocial attitudes depending on relative status. Only three subjects were competitive ($\rho < 0$ and $\sigma < 0$), and they displayed this pattern both towards their friends and the strangers, suggesting that they always tried to maximize their lot compared to the other's, regardless of the relationship (Figure 3.4(d)). Our parameter estimates for strangers are generally consistent with those found by Morishima et al, 2012.

Competitive and difference-averse subjects tended to have parameters of greater magnitude than the prosocial subjects, which may be driving the population averages to zero, or even negative numbers in the case of being worse off than strangers. This discrepancy between aggregate results and individual-level results suggests that looking at average population parameters may be misleading, and a more detailed look at the distribution of individual results may be called for (Blanco et al., 2011).

3.4 Conclusion

Here we conducted a systematic characterization of other-regarding preferences, with the other being a friend or a stranger. To isolate other-regarding preferences from reciprocity, which may exist independently of pure concern for others, we restricted ourselves to non-strategic decisions. In order to maximally distinguish between candidate models, we developed an efficient trial selection method, which we applied to a modified dictator game task. We used up to 150 trials per subject and found that, compared to a Rawlsian or Cobb-Douglas model, the Charness-Rabin model described human choices best. Examining the model parameters at the individual-subject level, we found that attitudes towards friends and strangers varied in degree more than in kind: subjects who were prosocial towards friends were prosocial towards strangers as well, albeit less so, and subjects who were competitive or difference-averse with strangers were so with friends as well. We found that in the neutral context of our experiment, prosocial attitudes, not difference-aversion or competitiveness, were most prevalent in the population: 10% of our sample were competitive, 10% were difference-averse, and the remaining 80% were prosocial. Prosocial subjects gave positive weights to others, friends and strangers, when better off and when worse off, albeit less so when worse off, and less towards strangers than friends. This across-the-board positive consideration of others does not imply that they are extraordinarily altruistic; some are rather close to indifference, especially towards strangers when worse off. However, prosocial individuals typically avoid investing in hurting others or pulling them down, regardless of relative status, akin to a “do no harm” rule. Taken together, our results show that people are responsive to relative status and relationship, and adjust their generosity accordingly, but rarely engage in antisocial attitudes solely for the reason of being worse off. This calls into question the emphasis on inequality-averse preferences, which require antisocial attitudes when worse off.

Our results are consistent with previous work showing that brain areas related to reward processing respond to gains for others as well as gains for oneself (Fareri et al., 2012; Waytz et al., 2012), and that the response to another person’s gain depends on our relationship to them (Strombach et al., 2015), with positive responses to gains for friends and negative responses to gains for antagonists (Braams et al., 2014; Singer et al., 2006). Our results contrast with those of Hutcherson et al's neurocomputational model of altruistic choices (Hutcherson et al., 2015), which found that generous choices were rare and concluded that they could be better understood as errors. However, that study used nine types of trials, all of them with a ($50, $50) default option, a set of trials which may not be the most conducive to finding moderate to low measures of altruism (see chapter 2 - An Information Efficient Trial Selection Method, Table 2.1). Perhaps more importantly, in that study the other was always a stranger. Including friends as recipients in our study was essential to finding a sizeable measure of generosity when better off. Thus, when investigating structures in the brain that may contribute to altruistic decision-making, including close relations might be the best way to ensure generosity is indeed elicited.

Our lack of evidence for pervasive difference-aversion preferences is in line with Blanco et al, 2010 finding that the Fehr-Schmidt model is a good descriptor of aggregate results, but not of individual data. This seems to be at odds with neural evidence presented by Tricomi et al, 2010. .In their study, a pair of subjects draw balls labeled “rich” and “poor”, the “rich” subject received a further payment of $50; then both proceeded to evaluated further monetary transfers to each of them or both in a fMRI scanner. Monetary amounts were symmetric, but they found that reward processing in ventral striatum and vmPFC, as well as subjective ratings of these transfers, differed in the two types of subjects. Specifically, they found that the contrast between transfers to Self and transfers to Other were greater in the “poor” subjects. However, strictly speaking, difference aversion requires negative weights for the other, not just reduced weights for the other compared to oneself or compared to “rich” subjects. Where they do observe negative weights for transfers to the other, this is observed only in the disadvantaged or “poor” subject and may be explained by the fact that they are looking at aggregate data. Moreover, studies aimed at measuring inequality-averse preferences such as Tricomi et al. 2010 tend to make the inequality variable very salient. This may cue participants as to what they are expected to feel about the upcoming stimuli and may explain that “rich” subjects valued transfers to the other more than transfers to themselves. By contrast, we strove to provide a neutral frame; our task design made the inequality theme minimally evident. We contend that this makes our results stronger, because it might induce subjects to express their true preferences rather than what they think they ‘ought’ to prefer. Of course, this does not imply that people cannot be influenced by framing. However, in an inequality-neutral frame, the antisocial aspect of difference aversion does not seem to come all that naturally.

Interpretation of results in terms of inequality-averse preferences seem to be more pervasive than inequality-averse preferences per se. For instance, a study conducted with pairs of children from 3 to 5 years old who received 2 and 4 stickers in return for cleaning up toys was titled: "Even 3 year-olds react to inequality" (LoBue et al., 2011). The data, however, showed that children reacted negatively to disadvantageous inequality, but not to advantageous inequality. Reacting against 'inequality' only when one is disadvantaged is not the most compelling argument for being inequality-averse. It is also worth noting that the same children who reacted negatively when they were passive recipients of a disadvantageous split may not be so inclined to sharing when they have to actively decide a split. We contend that active decisions are better tools to probe preferences than reaction to decisions made by others.

Our study has several limitations. (1) Model comparison is always limited by the preselection of models. It is possible that an untested model is still better than the Charness-Rabin model. (2) The trials selected as 'most informative' depend not only on the candidate models, but also on the pool of trials considered for selection. This preexisting pool of trials is somewhat arbitrary, and others may be considered. Furthermore, when calculating the 'informativeness' of a trial in future studies, one could use experimentally informed prior distributions over parameters instead of the uniform-on-an-interval distributions that we have used. (3) We have used the 'informativeness' of individual trials as a selection criterion. This could lead to redundant trials. Strictly speaking, mutual information is not a sum across trials. The theoretically best way to pick the N most informative trials is to calculate the mutual information for complete sets of N trials. For example, for characterizing Fehr-Schmidt agents, it is necessary to use both upward- and downward-facing trials. Our algorithm, however, selects only one upward facing trial. This is because the 24 downward facing trials individually are more informative than the upward-facing trials. However, the information is likely more redundant across downward-facing trials than between downward- and upward-facing trials. While in our design, this problem is alleviated because many trials selected for characterizing competitive agents face upward (see figure 3.4(b)), future work could attempt to take into account redundancy by estimating the information provided by sets of candidate trials, rather than individual ones. (4) Some authors have questioned the external validity of dictator games. In the classic dictator game – in which an endowment is to be split – subjects are susceptible to framing effects in view of the task being presented in terms of giving versus taking (Bardsley, 2008; List, 2007). Our two-alternative forced-choice task avoids both frames and may therefore be less sensitive to this concern. Others have questioned dictator game tasks due to violations of rationality when an unattractive lottery is introduced to the choice set (Oberholzer-Gee and Eichenberger, 2008) or due to marked drops in giving when the possibility of delegating responsibility is available (Dana et al., 2007). In both cases, the complexity of the decision scenario might introduce confounding variables (taking of responsibility may be costly even if not in pecuniary terms), produce violations of rationality, or introduce avoidance of extreme strategies (Kümmerli et al., 2010a). We have purposefully left out complex aspects of other-regarding preferences such as intentions (Falk et al., 2003; Rabin, 1993), need (LoBue et al., 2011), or merit (List, 2007) in order to focus on measurement, with the understanding that underlying trait dispositions maybe tapped into in this simple, neutral context (for reviews and meta-analyses, see Engel, 2011 and Camerer, 2011.

Our results do not purport to be an exhaustive or definitive description of decision-making involving others. After all, we have limited ourselves to non-strategic contexts, whereas real life is full of repeated interactions and reciprocity concerns. However, we believe that we have tapped into a natural disposition to take others into consideration, regardless of whether other motivations may contribute as well. In that regard, we can say several things: people tend to be more inclined to make sacrifices for people they feel close to (friends), and when they perceive there is a need for their generosity (their friend is worse off). Moreover, our results underline that altruism is not a binary variable, either present or absent, but a graded one. Many discussions motivated by the quest for the appropriate way to organize society are, at their core, posing a question about human nature: “are human beings selfish of altruistic?” Instead, it seems that people are altruistic to some extent, and selfish to another extent: most people will engage in sacrifices for others, but not just any level of sacrifice. A weight for friends of 0.304 with SD=0.363 on a scale from -1 to 1 implies that there are some sacrifices that people will make for their loved ones, but not others. Being aware of this range of altruism in the population could help set realistic expectations of others, both at the personal and the societal level (Gintis et al., 2008).

Chapter 4

The Effect of Cognitive Load on Other-Regarding Preferences

4.1 Introduction

A longstanding question regarding altruism is whether it comes naturally to us, or is it the product of self-control over selfish impulses? Must we thank millennia of culture and civilization for the taming of our base impulses, or are our impulses kinder than we give ourselves credit for? These questions tap into a centuries-old-debate concerning human nature, and related to that, a debate on nature versus nurture. In other words, how much of our behavior is a learned skill and how much of it is a natural disposition? Modern psychologists have made use of cognitive load tasks to try to answer this question. The rationale is that our capacity to exert cognitive control is limited, and by exhausting that capacity, we presumably let our impulses show. When cognitive load tasks are coupled with other-regarding decision tasks, we can hope to get some insight into the nature of altruism. If altruism is the product of self-control over a selfish nature, we should see that under cognitive load, our selfishness increases. On the opposite end, if altruism is a natural disposition, under cognitive load it should remain the same or increase.

The types of cognitive load tasks that have been used so far are varied, and as we will see in a review section further down, some of these tasks have proven to be more effective than others in taxing our cognitive control capacities. An additional limitation in this line of research is that the decision-making tasks used to measure altruism produce a somewhat coarse estimate of our other-regarding inclinations and compel researchers to rely on group averages. The results, therefore, are inconclusive. Here we aim to contribute to this effort by leveraging on the measurement precision obtained using the DGM trials (see chapters 2 - An Information Efficient Trial Selection Method T1 and section 2.1. Trial Selection 2.0 of this chapter). These information-efficient trials allowed us to have as precise parameter estimates as we could hope for at the individual level. As a result, in addition to the traditional hypotheses presented previously, we were able to consider the possibility of finding a significant degree of individual variability in the role of self-control on altruism.

4.1.1 History of a Debate

Since the earliest manifestations of philosophical thought, men have introspected about themselves, and by generalization, about human nature. Although the history of the ideas about this subject is fraught with political motivations, it is worth reviewing some of the most significant positions.

4.1.1.1 Born this way. Or, Naturally Noble or Nasty?

When we think about human nature, many of us will be reminded of the famously opposing views of Thomas Hobbes and Jean Jacques Rousseau. In the context of developing their political philosophies, they speculated about men's 'state of nature', before institutions, 'civilization', and social norms. What was in our nature was crucial in designing the proper institutions and governing bodies. Understandably, it was paramount to know what 'we are working with'.

Hobbes envisioned the state of nature as a permanent war of every man against each other. Without a political authority to keep people in check, men lived in "continual fear, and danger of violent death". His picture was so bleak, that his description of life in the state of nature became exceptionally famous: "and the life of man, solitary, poor, nasty, brutish, and short" (Hobbes, 1651). His view of men was evidently on the selfish and vicious side of things.

Rousseau espoused the polar opposite view. He thought that in the state of nature, men lived peacefully and untroubled. Although still lacking in high virtues such as love and loyalty, art and industry, pre-state humans were also devoid of the greatest sins. Naturally inclined to self-preservation and compassion, "nothing can be more gentle than him in his primitive state" (Rousseau, 1755). To Rousseau, wickedness and corruption came from society: "There is no original perversity in the human heart. There is not a single vice to be found in it of which it cannot be said how and whence it entered" (Rousseau, 1762). Although he did not coin the term 'Noble Savage', his philosophy came to be viewed as the standard-barer of this idea.

With their obvious differences, both philosophers relied on the idea of an innate disposition within us. They were not, however, biological determinists. They thought society had a significant role in ultimately shaping our behavior. In any case, if Hobbes was correct, then altruism is the result of the taming effect of civilization, and so we should expect to see that cognitive load reduces our pro-social inclinations. If, on the other hand, Rousseau had it right, civilization is to a great degree a degenerating force that corrupts people's hearts, and therefore, we should see that under cognitive load, our pro-social inclinations are unleashed.

4.1.1.2 It's the culture, stupid! Or, the theory of the Blank Slate

In stark opposition to the idea of any innate characteristic, and even innate ideas, is the concept of the 'Blank Slate'. It is attributed originally to John Locke, although he used a different metaphor. In *An Essay Concerning Human Understanding*, he writes:

> "Let us then suppose the mind to be, as we say, a white paper void of all characters, without any ideas. How comes it to be furnished? (...) Whence has it [come] all the materials of reason and knowledge? To this I answer, in one word, from experience" (Locke, 1690)

During his time, this idea served to challenge dogmatic justifications of the status quo: the divine right of kings, the authority of the Church, and even the institution of slavery, as no one could claim innate merit or superiority over others. Locke's intellectual heir, John Stuart Mill, used an equivalent reasoning to argue for the equality of women and against racism.

> "I have long felt that the prevailing tendency to regard all marked distinctions

> of human character as innate, and in the main indelible, and to ignore the irresistible proofs that by far the greater part of those differences, whether between individuals, races, or the sexes, are (...) produced by differences in circumstances, is one of the chief hindrances to the rational treatment of the great social questions, and one of the greatest stumbling blocks to human improvement." (as quoted by Fox, 1989)

Here we begin to see that the concern is no longer necessarily a universal human nature but differences between groups and the potential pseudoscientific justifications for racism and sexism. One can sympathize with the fear that the concept of human nature may erroneously lend credence to the concept of a female or black nature, used to justify their oppression. But the reaction might be misdirected. Whatever human nature there might be, by definition it would have to apply to all humans, males as well as females, of whatever skin color. The concept of human nature, by definition, appeals to what is common to all of mankind.

We also start to see an association between innate and 'indelible', to use Mill's words. This association is lacking in previous philosophers, as we saw in the case of Hobbes and Rousseau, and is also lacking in modern instantiations of the premise of innate characteristics, such as in behavioral genetics. Modern genetics allows us to understand the influence of genes as probabilistic and in interaction with the environment. However, again, in their time the fear was, understandably, of a petrifying determinism.

In the field of psychology, the Blank Slate position was taken to an extreme by the founder of behaviorism, John B. Watson:

> "Give me a dozen healthy infants, well-formed, and my own specified world

> to bring them up in and I'll guarantee to take any one at random and train him to become any type of specialist I might select – doctor, lawyer, artist, merchant-chief, and yes, even a beggar-man and thief, regardless of his talents, penchants, tendencies, abilities, vocations, and race of his ancestors" (Watson, 1924)

Both Mill and Watson's quotes were followed by what could be construed as confessions of political motivation that question the extent to which they truly believed in the Blank Slate position. In the case of Mill:

> "[This tendency] is so agreeable to human indolence, as well as to conservative interests generally, that unless attacked at the very root, it is sure to be carried to even a greater length that is really justified by the more moderate forms of intuitional philosophy" (as quoted in Fox, 1989).

These words suggest that Mill considered the belief in some level of human nature justified but had to be opposed 'at its very root' nonetheless. In the case of Watson:

> "I am going beyond my facts and I admit it, but so have the advocates of the contrary and they have been doing it for many thousands of years" (Watson, 1924)

The motivation to fight the possible implications of certain ideas is likely behind variations of the Blank Slate position that we find later on. During the 20th century, the most prominent and successful expression of this position is what came to be known as Social Constructionism. The term was introduced in 1966 by sociologists Peter Berger and Thomas Luckmann with the publication of their book, *The Social Construction of Reality.*

This philosophy contends that things we take for granted are not natural at all, but actually constructed by society, and therefore artificial. In his book, *The Social Construction of What*, philosopher Ian Hacking provides a list of titles that followed, including *The Social Construction of*: *Emotions* (Harré, 1986), *Facts* (Latour and Woolgar, 1979), *Nature* (Eder, 1996), and *Danger* (McCormick, 1995), to name a few that may shock most people. Other less shocking titles develop the idea of gender as a social construction (Dewar, 1986; Lorber and Farrel, 1991), although there are strong cases for natural sex differences - on average (Brizendine, 2006; Buss, 1989; Croson and Gneezy, 2009; Feingold, 1994).

As it picked up in the social sciences, constructionist thinkers tended to assign an intentionally oppressive role to the 'construction'. In view of that, Hacking suggests that the popularity of this philosophy is due to its potentially liberating effect: if the way things are is not natural and do not need to be this way, we can change them to fit our ideals. For more on this, see *The Social Construction of What*, Ian Hacking 1999, and *Blank Slate*, by Steven Pinker, 2002.

In any case, if the Blank Slate theory is correct, and we are entirely socially constructed, we should see that cognitive load has no effect on our pro-social inclinations. Since there is no innate tendency over which we are exerting self-control, depleting our control capacity should be inconsequential.

4.1.1.3 Not Darwin Again...

Yes. Incredibly enough, Darwin speculated that perception, cognition, and emotions, very much like physical traits, had evolved as biological adaptations (Darwin, 1872). This insight inspired the ‘father of American psychology’, William James, who went as far as suggesting that psychology should be a division of biology. He was prominent in what came to be known as functional psychology, which stressed that mental processes served specific functions during our long evolutionary existence. His theory gave a central role to instincts, and even speculated that humans had more of those than animals!(Buss, 2015). William’s brand of psychology, however, was relegated to the sidelines during the apogee of behaviorism.

Later in the 20th century, the naturalist viewpoint resurfaced with the publication of E.O.Wilson’s *Sociobiology: The New Synthesis* (1975). By the author’s definition, sociobiology is the ‘systematic study of the biological basis of all social behavior’. As such, it understands behavior as arising from genes and shaped by natural selection. The book was primarily a work of behavioral ecology, a compilation of theoretical developments and empirical studies of animal social behavior, mainly focused on non-human animals. However, Wilson explicitly included examples from our own species, the Homo Sapiens, and devoted the last chapter exclusively to speculate about us.

What followed was a virulent controversy. According to Finnish sociologist Ullica Segerstrale, who wrote an illuminating account of the debacle, the initial reception of Wilson’s work was positive, with many of his fellow biologists ‘grateful’ to him for processing such a massive amount of literature ‘for’ them. Some contended that he had not said much new but recognized his contribution as having created a field ‘by showing its scattered practitioners that it existed’. This was however after the horrors of the Holocaust, and the social climate was suspicious of evolution and the genetics of behavior, especially when it came to humans. Moreover, at his time the prevailing view in academia was that culture, or the environment, was the main explanation of human behavior.

Soon after the publication of Wilson's book, his detractors organized in what eventually came to be the Sociobiology Study Group of Science for the People. It consisted of two of his Harvard colleagues, fellow evolutionary biologists Richard C. Lewontin and Stephen J. Gould, and others from different departments in a variety of academic positions[1]. The most polite accused Wilson of engaging in 'bad science', or of excessive 'biological determinism', despite Wilson himself clarifying that he considered the environment to be an important component of human behavior and warning against deriving a moral philosophy from biology. The least polite once stormed his speaker podium at a conference chanting 'Racist Wilson you can't hide, we charge you with genocide!' and poured a jug of ice-water over his head[2]. In Segerstrale's final veredict, the controversy was to a large degree manufactured, but sincerely fought on both sides. For more on the Sociobiology debate, see *Defenders of the Truth*, Ullica Segerstrale, 2000.

Interestingly, Wilson shared with his detractors the belief in the ultimate coupling of science and moral values. For this reason, in developing sociobiology, Wilson made the question of 'altruism' absolutely central. In the book, he presented several evolutionary theories of altruism, including the now widely accepted theories of kin selection, proposed by Hamilton (1964), and reciprocal altruism, proposed by Trivers (1971). He even gave great consideration to the group selection theory of altruism, which was strongly questioned by other sociobiologists such as Richard Dawkins. Group selection might be the strongest case for 'real' altruism, as it does not imply hidden benefits either in the future or to relatives carrying the same genes. Dawkins, keen of separating science from morals,

[1] A pre-medical student, a graduate student, a research assistant, a teacher, a resident fellow, a research associate, a doctor, a psychiatrist, and professors in biology, anthropology, zoology, and at the medical school (Segerstråle, 2000)

[2] Wilson proceeded with his presentation, soaking wet but calm, and received a standing ovation afterwards (Segerstråle, 2000)

dismisses the group selection theory of altruism, and understands reciprocal altruism and kin selection ultimately as expressions of selfish genes. And so, in his famous book The Selfish Gene, he urges: ‘Let us try to teach generosity and altruism, because we are born selfish” (Dawkins, 1989)

After the commotion around sociobiology subsided, the naturalist viewpoint was most successfully articulated in a new discipline, appropriately named evolutionary psychology. More focused on psychological processes than on behavior, evolutionary psychology shares the same principles as sociobiology: the idea that our mind, as any other part of us, has evolved to solve problems that we have faced during our ancestral past, by natural selection. It draws from evolutionary biology, anthropology, and psychology, and it looks for universal explanations for universal questions. Cross-cultural consistency is a staple of this field (Barkow et al., 1992; Ekman and Davidson, 1994).

This evolutionary instantiation of the naturalist viewpoint builds up from the empirical observation of nature and strives to find evolutionary explanations of the observed phenomena. In the case of social behavior, they take for granted that altruism (or altruistic behavior) exists, and then proceed to understand how that could be, from an evolutionary perspective. This does not constitute a denial that cultural norms may operate on top of natural dispositions. In terms of what the naturalist view implies for the effect of cognitive load on social preferences, in so far as they recognize that some level of altruism is natural and culture acts as a modifier on that, its effect could go on any direction. It will depend on whether each individual is culturally trying to enhance or suppress their instinctual reaction.

4.1.2 Literature Review

In this chapter, we address the question of whether altruism is an impulse or the product of self-control through the prism of cognitive psychology. The key process here is cognitive control, a top down regulatory mechanism that may dampen (or amplify) bottom-up emotional impulses. Cognitive control is understood to be limited, a resource that can be depleted or cut short through several methods. In this section we will describe the various ways experimental psychologists have manipulated cognitive control and tested its role on social decision making.

4.1.2.1 Types of Cognitive Load

- Time Constraint

  Cognitive control over emotional impulses take time to take effect. One popular way of limiting cognitive control has thus been implementing time constraints. The presumption is that if people do not have time to engage in deliberative processes, their impulsive behavior takes over. Experimenters using this technique have pressured participants to make decisions fast. There are several limitations to this technique. One is that what is considered fast by one experimenter is slow for another. Is 10 seconds to make a dictator game choice too fast? Another limitation of this technique is the problem or reverse inference (Poldrack, 2006). We will see this argument in concrete terms when we recount the experimental results and the discussion that ensues between researchers in favor and against its use.

- Digit Span

  Another popular technique to limit deliberative capacity is to require participants to remembers a string of digits while they are performing the main experimental task. The string can consist of solely numbers, or combinations of numbers and letters. The control condition may be a string of equal length but of easier memorization, such as a number with a structure (1234567 or 7777777), or a shorter string, or no memory requirement at all. Typically, the string is presented to the participant at the beginning of the task, and they are asked to reproduce it at the very end. The limitations of this technique are several. One is that the cognitive load can be argued to be too low. Once the string has been encoded, the participants can forget about the issue until the very end, putting in question whether it is actually taxing their cognitive capacity during the main experimental task. In fact, when we review the experimental results, we will see that many experimenters recognize this fact. For the most part, this method has been abandoned for a lack of results.

- *N*-Back

  This technique produces a stream of stimuli that must be constantly updated, providing a much more challenging cognitive load than any of the other methods. On every trial, a different stimulus is presented. It can be a letter or a number. A hard condition requires the participants to remember the stimuli presented in the current trial and the previous N trials. They are asked to respond whether the current stimulus is the same as N trials back. The bigger the N, the greater the difficulty. 2-back is a common choice for a hard condition. On a 2-back condition, on every trial three stimuli must be held in memory, compared, and updated. On an easy condition, a smaller number of stimuli must be held in memory. In the easiest case, a 0-back, only the current stimulus must be processed. This method imposed a cognitive load during the entirety of the main experimental task.

4.1.2.2 Experimental Results

One of the most impactful studies on whether altruism is intuitive or deliberative was conducted by Rand, Greene, and Nowak in 2012, hereby referred to as RGN (Rand et al., 2012). This study generated a series of responses and counter responses and for this reason we address it first. RGN pools a series of social dilemma studies, such as the Public Goods Game (PGG) and the Prisoner's Dilemma (PD), conducted both online through Amazon Turk, tapping onto the general population, and in person by recruiting students to their lab. In the cases where there were repeated interactions, they considered only the first move. They first looked at response times (RT). The faster half of the responses showed greater levels of contributions than the slowest half. To explore the causal link of intuition on cooperation, they manipulated the participants' decision times by putting them under time pressure or imposing a time delay when recording their choices (under or over 10 seconds respectively). Again, contributions were higher under time pressure compared to time delay. Finally, they tested the effect of a priming manipulation that induced relying on intuition versus reflection. They found that when they promoted intuition, contributions where higher than when they promoted reflection.

Tinghog et al 2013 immediately noticed that RGN excluded a significant portion of responses that failed to comply with the time restriction. Because slow responders tended to cooperate less as per RGN results, excluding late responders necessarily skewed results towards more cooperative faster choices. They also noticed that RGN instructions might have primed subjects in their hypothesized direction. They conducted several studies aimed at addressing these concerns and found no effect of time pressure on cooperation. This challenging result adds to the finding by Piovesan and Wengstrom, 2009, where they observe that fast responders in a dictator game tended to be more selfish than slower ones. Rand et al, 2013 fired back at Tinghog et al, 2013. They reanalyzed the data following Tinghog's group suggestions and confirmed their original effect of time pressure on cooperation. In their response, they also noted that the original RGN paper considers 10 studies, encompassing three different methods, of which time pressure is only one.

On the following year, Rand and his colleagues published a paper where they expanded their theory of intuitive cooperation, offering an explanation for the variability of results across studies (Rand et al., 2014). They proposed the Social Heuristics Hypothesis (SHH), which predicts when and for whom intuition will boost cooperation. According to the SHH, people internalize strategies that work for them in real life. These may be social norms that have become automatic. They then bring those internalized strategies with them into atypical situations, such as laboratory experiments. Deliberation is thus required to override these generalized automatic responses in order to adapt behavior to the specific context. The SHH predicts that intuition will favor cooperation only in people who i- come from environments where cooperation is favored, and ii- have little experience with laboratory experiments. People who come from environments where cooperation is not supported will have internalized defection as a default strategy. For them, there will be no conflict between intuitive and deliberative processes (deliberation here is assumed to favor selfishness), and so, time pressure will have no effect on their cooperation levels. Likewise, those who have extensive experience with laboratory experiments will have developed new automatic responses appropriate to the 'atypical' setting. Unlike naïve subjects, experienced participants will not be bringing their daily lives intuition into the lab. They will experience significantly less conflict between intuitive and deliberative processes and will show significantly less or no effect of time pressure on cooperation. And in fact, they found evidence supporting these hypotheses. They compared Amazon Turk contributions as it became more popular, from February 2011 to January 2013. Here time is a proxy for experience. They found that contributions under time pressure decreased as Amazon Turk became popular, but they did not observe the same effect under time delay. They also run another MTurk experiment that specifically asked participants about their previous experience. Naïve subjects showed a greater level of contributions under time pressure compared to time delay, but no significant difference was found amongst experienced subjects.

Perhaps the strongest critique to Rand's group results was articulated by Krajbich et al, 2015. They question the practice of 'reverse inference' from response times (RT) because it does not take into account other possible sources of variability in the data (Poldrack, 2006). In this case, they point to strength of preference. Choices between options similar in subjective value will take longer, whereas choices between a clearly superior and a clearly inferior option will be much faster. The same happens in perceptual decision making between similar stimuli that are hard to distinguish. Krajbich's group demonstrate their point by running the same RGN task, plus a dictator game and an intertemporal choice task, but with several settings in terms of costs and benefits of cooperating, giving, and waiting. By changing these settings, they manipulated the strength of preference of one choice over the other. As a result, they were able to replicate, eliminate, and reverse the negative correlation between RT and selfishness found by RGN.

While the RT 'reverse inference' critique is unquestionably challenging to the intuitive altruism hypothesis, it leaves other RGN measures unaddressed, such as time restrictions and intuitive priming. Another group of papers relegate time measures or drop them altogether and make use of cognitive load tasks instead. These tasks purport to drain participants of their deliberative capacity, presumably bringing up our intuitive responses. A key issue when considering these kinds of tasks is whether they were indeed effective in draining subjects of cognitive control capacity.

Cappelletti et al, 2011 used a 3-digit cognitive load task on an Ultimatum Game (UG) and looked at the proposers' offers. In the control condition, participants were not requested to memorize any number. They did not find an effect of cognitive load on ultimatum offers. One may think that 3-digits is too weak of a load, if at all. They also ran a time pressure version of this task and found a positive effect on offers. Cornelissen et al, 2011 tested a 7-digit cognitive load task on a classic 1-shot dictator game task (DG). Before the DG decision, subjects were given either a random (hard) or structured (easy) 7-digit number that they should remember after the DG task was over. Participants were classified as prosocial or proself using the Social Value Orientation scale. They found that cognitive load increased altruism in prosocials, but no in proselfs, a result consistent with RGN's Social Heuristic Hypothesis. Hauge et al, 2016 ran a 7-digit cognitive load task on a series of 1-shot DG questions. Again, the easy condition consisted of structured digits, while the hard condition consisted of random ones. They found no effect on generosity but recognized that the cognitive load task may not have accomplished the goal of depleting subjects' control capacity. Benjamin et al, 2013 used 7-digits versus nothing to evaluate the effect of cognitive load on risk and time preferences of Chilean high-schoolers. They found no effect. Only Shiv and Fedorikhin, 1999 found an effect of 7-digits versus 2-digits, but they literally had to put a chocolate cake under participants noses. The choice was between the cake and fruit salad, the former being considered impulsive, and the latter deliberative. But if the choice was symbolic (by checking a piece of paper without seeing the options), the effect disappeared.

As a whole, the digit version of the cognitive load task does not seem to be very effective. Schulz et al, 2014 introduced an auditory N-back task in simultaneous to a series of 20 mini dictator games. Participants heard a letter at the beginning of each trial, before proceeding to make their DG decision. In the hard condition, 2-back, they had to press a key if the letter was the same as 2 trials back. This type of task requires constant monitoring and updating and constitutes a much more significant load than remembering a 7-digit number at the very beginning of the experimental session to be retrieved at the very end. In the easy condition, 0-back, participants had to indicate when they heard the letter "L". For every mini DG with an equal split they included one with an almost equal split. This allowed them to test the hypothesis that people default to the equality heuristic under cognitive load. They found an increase in the proportion of 'fair' choices under high cognitive load, but it was not explained by the equality heuristic, as this tendency did not differ in games with an almost equal, as opposed to an equal, option. Additionally, subjects under the low cognitive load condition were more responsive to the incentives presented by the task. That is, the larger the inequality in the 'unfair' option, the greater the likelihood of choosing the 'fair' option. This was not the case for subjects under high cognitive load, who chose the 'fair' option at a consistently high frequency, regardless of the level of inequality in the game.

4.2 Methods

Here we apply an N-back cognitive load task in simultaneous to a 2AFC Dictator Game task. As opposed to previous studies, where the main variable is a rather coarse measure of altruism -the percentage of 'fair' choices or total contribution, averaged across participants in a between-subjects design, - we will be looking at individual level effects of cognitive load on other-regarding preference parameters. To do this, we leverage on our ‘information maximizing’ DGM trials. As with our other studies, we look at decisions made towards friends and strangers.

4.2.1 Subjects

Participants were recruited from the general population through advertisement on Craigslist. 26 people registered. They were asked to come with a friend, so in total 52 people came to the Center for Experimental Social Science (CESS) Lab, where the experiment was conducted. Friends participated as decision makers as well.

We ran 6 sessions of between 4 and 16 people each. The number of people on each session was a multiple of 4, as each pair of friends were matched with another pair that would act as their strangers for the decision-making purposes.

When they arrived at the lab, they were welcomed by the experimenter, handed a packet of paperwork that included the consent form and instructions, and directed to take a seat at one of the available computer stations. They waited until everyone arrived. In sessions where the number of people was not a multiple of 4, the last 2 to arrive where paid the show-up fee and dismissed. Once everybody was ready, the experimenter delivered verbal instructions and answered questions.

4.2.2 The N-back task on the Dictator Game

The dictator game task was exactly the same as the one used in our previous experiment (see chapter 3 - A Systematic Characterization of Other-Regarding Preferences), except for the specific amounts selected for each trial. Each trial consisted of 2 options; and each option, of two amounts: one for self and one for other. The other was a friend in half of the trials, and a stranger in the other half. The task was self-paced, with a maximum of 10 seconds per choice, which is typically plenty of time.

In the cognitive load version of the dictator game, a letter was presented in the center of the screen for 0.5 seconds at the beginning of each trial. We limited the letters to vowels, to keep the difficulty of the task to manageable levels. In the 'Hard' condition, subjects had to press a key indicating whether the letter on the current trial was the same as two trials back, before making their dictator decision. This required constant monitoring and updating of the last two letters and comparison with the current one. In the 'Easy' condition, they had to indicate whether the letter in the current trial was an 'A'. Every trial required a key press regarding the letter, either for 'yes' (it is the same as two back or it is the letter 'A') or 'no'. If they did not respond to the letter task before making their dictator decision, that trial was recorded as incorrect for the purposes of the n-back task. The letter sequences were random but modified to assure a 20% of 'yes' responses were required in every block.

Each participant responded to 300 trials in total. This arises from 75 selected trials in 4 conditions: hard and easy, with friend and stranger (Hard-Friend, Easy-Friend, Hard-Stranger, Easy-Stranger). Trials were blocked in sets of 25. Hard and Easy blocks were stringed together, and friend and stranger blocks were stringed together as well, to avoid excessive confusion among participants. The order of these conditions and of the trials within each block, however, was randomized.

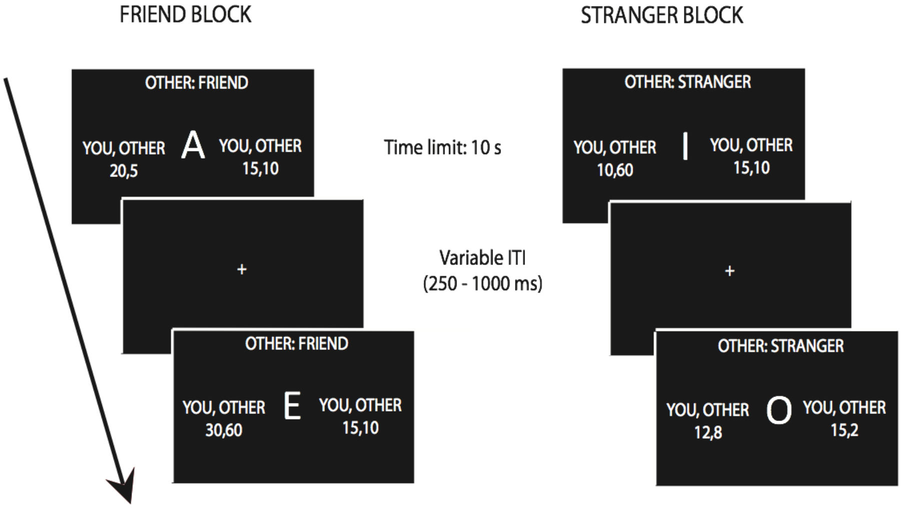

Figure 4.1 Cognitive Load Task. N-back task in simultaneous to a 2AFC Dictator Game (DG). A vowel appears in the center of the screen for half a second. Participants are required to respond 'yes' or 'no' (is it the same vowel as N trials back) before making their choice on the DG

4.2.3 Trial Selection 2.0

The amounts on each option were selected using our ‘information maximizing’ trial selection method. We started developing this method for an earlier experiment and continued improving it for this one. In this last stage of the method development, we made two improvements. First, we used data collected in previous studies to incorporate priors over parameters. Our original method assumed uniform priors over models first, and once we knew which model worked best, over parameters (see chapter 2 - An Information Efficient Trial Selection Method). We know from previous experiments that the Charness-Rabin model describes the vast majority of the population best, and within that model, we have a pretty good idea of what the distribution of parameters is (see chapter 3 - A Systematic Characterization of Other-Regarding Preferences). We used that information as priors, so that the trial selection method specifically targets areas of the parameters’ space that we know are more densely populated.

The second improvement is the consideration of dependence of the information provided by different trials. In the original version of our trial selection method, we made the simplifying assumption that trials were independent in the information they provided. But responses given to any set of trials by the same individual will not be independent. Therefore, the information they provide may be overlapping. In other words, the third most informative trial may be very informative when standing alone, but not when I have the information provided by the second most informative trial. As we have noted before (see chapter 2 - An Information Efficient Trial Selection Method), the correct way to choose N most informative trials is to consider every possible combination of N trials. But this becomes computationally untenable very fast.

We solved this issue by iteratively considering an ever-increasing set of trials. In the first round, we evaluated each trial separately (independently), and selected the single most informative one. We added that to a ‘history’ of responses and removed it from the pool of remaining trials. On the following iteration, we again considered each trial separately, but including the response given in the history trial. On this second round, the selected trial would be the second most informative one considering dependence. On each iteration, we selected the single most informative trial, added that trial to the history, and removed it from the pool of remaining trials. In this way, we were able to consider dependence of information but keeping the computational demand to feasible levels.

We repeated this iterative process with 20, 50, and 75 trials. We used the smaller set of trials to check the consistency of the method on bigger sets. For instance, the first 20 trials of the set of 50 were mostly the same as the 20 trials of the set of 20. We considered this method successful and, based on highly demanding standards for parameter recovery performance, we chose the set of 75 trials to move forward with the experimental task.

4.2.4 Incentive Compatibility

The dictator game task was incentivized in the same way as in our previous experiments. At the end of the session, one trial was selected at random and the choice on that trial was implemented for payment. If the participant had missed that trial (did not record a dictator response on time) there would be no extra payment to the show-up fee. No participant had this problem.

The cognitive load task was incentivized in the following way. Performance was recorded separately for the hard and easy blocks. The trial selected for payment could come from either of those conditions. The payment amounts where multiplied by the performance factor, which was a number between 0 and 1. So the better the subject performed, the least their payment would be adjusted down. This proportional payment maintains the incentives intact in terms of revealing their true preferences in the DG task.

4.2.5 Excluded Subjects

Three participants did not finish the DG task and were excluded from the analysis. Two of them were excluded due to poor comprehension of the English language and lack of experience interacting with computers. The third had a motor disability and started to feel pain on his fingers.

Two additional subjects were excluded from further analysis upon realizing that they had a performance equal to zero in one of the conditions of the N-back task. It is likely that they got distracted and forgot about this aspect of the experiment, as it would be difficult not to get any answer correctly at least by chance. One of them got a performance of zero on the easy condition, which only required to state when they saw the letter A on the screen. This is even easier than registering for the experiment, so it reinforces the suspicion that they simply forgot about the N-back task.

4.2.6 Model fitting and bootstrapped s.e.m for individual parameter estimates

We fitted the Charness-Rabin model at the individual level for each of the 4 conditions using the BADS algorithm developed by Acerbi and Ma (Acerbi and Ma, 2017). BADS, for Bayesian Adaptive Direct Search, is a hybrid Bayesian Optimization method that incorporates elements of MADS, the classic Mesh Adaptive Direct Search. MADS attempts to improve an optimization solution by testing points in the vicinity of the current point under evaluation, iteratively moving towards a better solution. BADS takes this feature of MADS and combines it with Bayesian optimization. BADS is recommended for large parameter spaces with no closed-form likelihood functions, which is not our case. In retrospect, perhaps the built-in matlab function fmincon would have been faster, producing equally satisfactory estimates.

In any case, evaluating the effect of cognitive load at the individual level was challenging because we were working with point estimates. One cannot perform statistical tests on point estimates. In order to overcome this challenge, we resorted to bootstrapping. This allowed us to obtain standard errors of the mean for every point estimate. We drew 75 random trials with replacement 100 times for each individual and each condition and used those bootstrapped data sets to perform the model fitting. As a result, we obtained 100 observations for each point estimate of interest. We then used t-tests on the bootstrapped estimates to evaluate the effect of cognitive load on individual level parameters.

4.3 Results

4.3.1 N-back Performance

Every participant saw a diminishment of their performance on the hard condition compared to the easy one. On average, performance on the easy condition was of 0.9763 ± 0.0043, while performance on the hard condition was of 0.816 ± 0.017. A paired t-test revealed a significant difference between these two conditions ($t(46)$ = -9.74, $p <$ 9.25e-13). We concluded that, as a group, the N-back task indeed taxed the participants' cognitive capacities.

4.3.2 Parameter Estimates and the Effect of Cognitive Load

We fitted a Charness-Rabin model to participants' choices on the 4 experimental conditions: Easy-Friend, Hard-Friend, Easy-Stranger, Hard-Stranger. For each condition, we obtained a decision noise parameter and two other-regarding parameters: weight for other when better off (ρ) and weight for other when worse off (σ). Because we intended to perform an individual level analysis, each fitting process involved bootstrapping 100 times. The results we present, therefore, consist of the mean and standard error of the mean (s.e.m.) of the bootstrapped parameter estimates for each participant, for each condition.

Figure 4.2 shows the parameters estimates (with s.e.m.), separately for friends and strangers (panels **(a)** and **(b)** respectively). In dark colors are the parameters in the easy condition, which constitute our baseline. In lighter colors are the parameters in the hard condition, which represent decision-making under cognitive load. A grey line connects parameters corresponding to the same subject in the two conditions. We can see that the distribution of these parameter estimates corresponds to what we have seen in previous experiments and other published studies (see chapter 3 - A Systematic Characterization of Other-Regarding Preferences, and Morishima et al, 2012). That is, most participants lied in the prosocial quadrant, below the 45-degree line, indicating that they had positive weights for others both when better off and worse off, but lower in magnitude when worse off. A smaller number of participants lied in the difference-averse quadrant, with a positive weight for others when better off, and a negative weight for others when worse off. And very few participants lied in the competitive quadrant, where regardless of the situation, their weight for others was negative.

In accordance with the results on chapter chapter 3 - A Systematic Characterization of Other-Regarding Preferences, 37 subjects out of the final sample of 47 had the same type of preferences towards friends and strangers (lied on the same quadrant) on the baseline condition, and 33 showed the same pattern under cognitive load. 26 subjects displayed the same type of preference on every condition, towards friends and strangers, both under cognitive load and on the baseline condition. All of the 26 subjects who maintained the same type of preferences in every context were prosocial.

In regard to the effect of cognitive load on each individual, we determined that there was a significant effect if at least one the two parameters was statistically different under the baseline and cognitive load conditions, as assessed by a paired sample t-test of the bootstrapped parameter estimates. Significance levels were evaluated using the Holm-

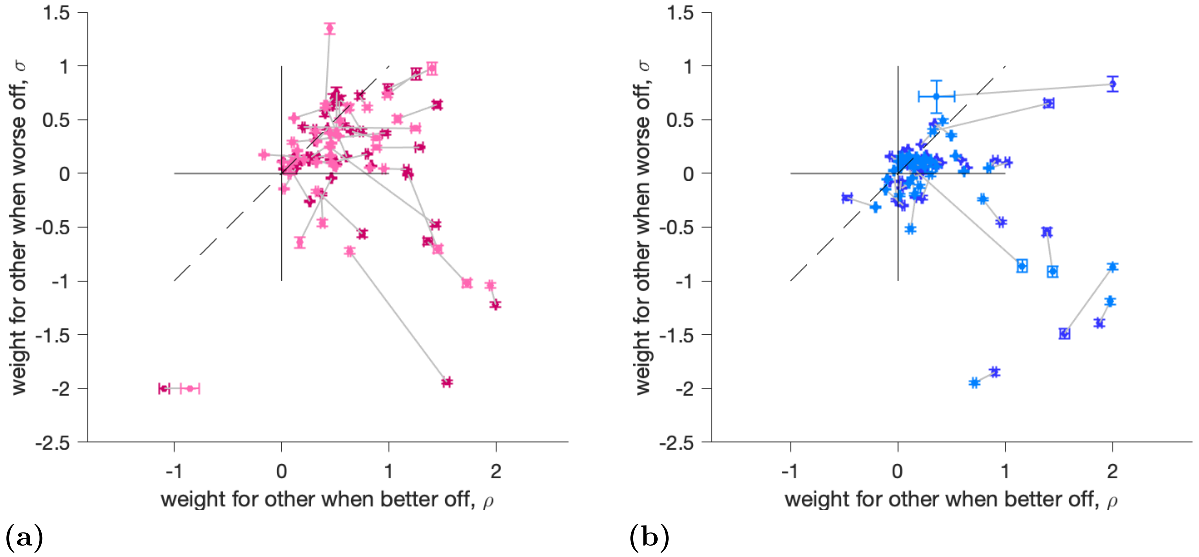

Figure 4.2 Other-Regarding Preferences, with and without Cognitive Load. Parameter estimates on the baseline condition (dark hue) and under cognitive load (light hue) for decisions involving friends **(a)**, and strangers **(b)**.

Bonferroni correction for 188 multiple comparisons (47 subjects*2 parameters*2partners). 42 out of the 47 subjects showed a significant effect of cognitive load in at least 1 of the 2 parameters in decisions involving friends, and 42 subjects showed this pattern in decisions involving strangers. There was no overlap of subjects showing no effect of cognitive load when it came to friends and strangers, meaning that only 37 of the 47 showed a significant effect of cognitive load on both friends and strangers.

In terms of the direction of change, which indicates whether cognitive load made people more prosocial, more competitive, or more or less difference-averse, there was enormous individual variability. Figure 4.3 shows the magnitude and the direction of change (with s.e.m.) separately for friends and strangers (panels **(a)** and **(b)** respectively). In this figure, the origin represents the parameter combination (ρ,σ) for each individual on the baseline condition. When it came to strangers, the magnitude of change was generally smaller ($t(46) = 2.201$, p= 0.016, paired, tailed. not tailed: p= 0.033), which is reflected in the more compact look of panel **(b)**. Of the 37 subjects who showed a significant effect of cognitive load for friend and stranger, almost half, 16, expressed the change in the same direction with both partners. The direction of change per se was well distributed in all directions, both for the 16 subjects who changed in the same direction with friends and strangers, as for those who changed in different direction with different partners. Even for the 26 subjects who remained in the prosocial quadrant in all conditions, the direction of change was well distributed in all directions.

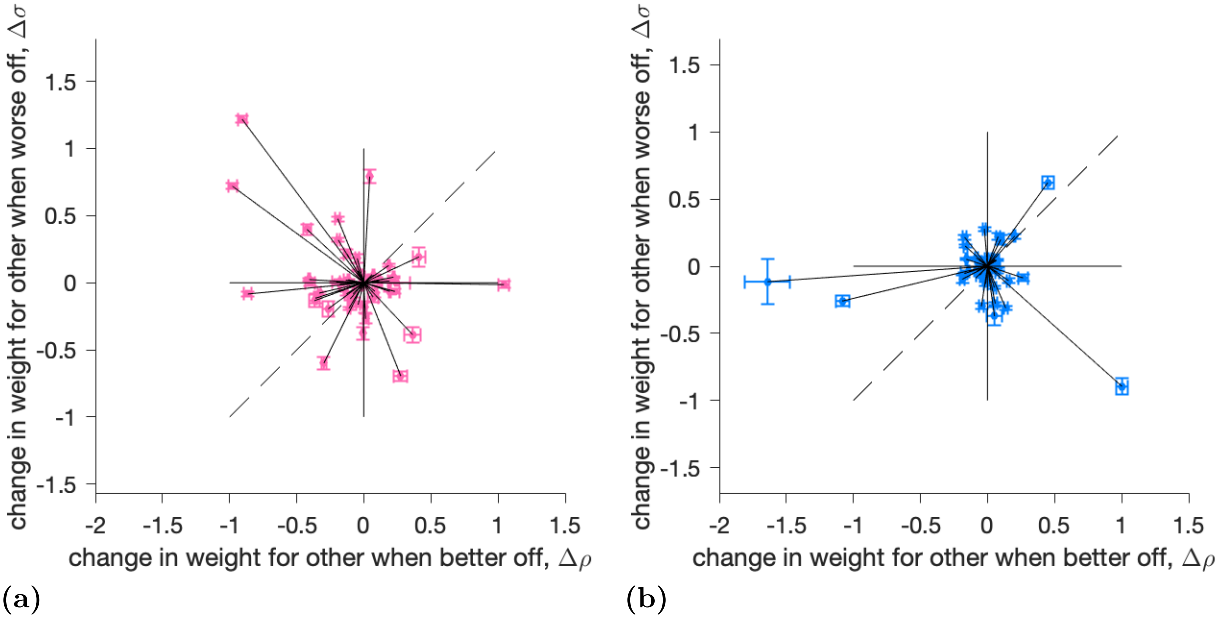

Figure 4.3 Effect of Cognitive Load. Change in other-regarding parameters under cognitive load compared to baseline, for decisions involving friends **(a)**, and strangers **(b)**. Error bars represent bootstrapped s.e.m.

Overall, there is a great amount of individual variability in the effect of cognitive load. This suggests that people have different guiding principles when they attempt to modify their behavior. We therefore cannot say that i- society corrupts our noble nature, nor that ii- it ennobles our base instincts. We can say, simply, that i- there is a cognitive contribution to social decision-making and that ii- it operates on top of our natural dispositions, which are unique to each one of us, and quite variable.

4.4 Conclusion

In this chapter we addressed a long-standing question regarding the nature of altruistic preferences: are they the product of self-control over selfish impulses, or on the contrary, are they an impulse in and of itself? We started by presenting different philosophical positions, from early philosophers centuries back, to modern scientists on this century and the previous one, and elaborated on the discussion between them. The debate at times got heated up, with closely held moral values with political implications at stake. Those determined to separate science from morals watched the carnage from the sidelines. In brief, we presented the following positions: 1) there is a human nature and it is nasty, primarily represented by Hobbes, 2) there is a human nature and it is noble, primarily represented by Rousseau, 3) there is no human nature, it is all a social construction, represented by Locke, Mill, Watson, and the social constructionist philosophy, and 4) there is a human nature and it is shaped by evolution, represented by Wilson and evolutionary psychology.

It is unclear what prediction the social constructionist view would make of constraining self-control, as in their view we do not seem to have impulses (instincts) to begin with. For the other three, the prediction depends of what they think of human nature. Hobbes would say that under limited self-control, we would revert to utter selfishness. Rousseau would probably say that our altruism would be unleashed. The evolutionary psychology crew would probably predict a variable effect depending on context, and some of them would agree with Hobbes.

We then reviewed how modern cognitive scientist attempted to address this question, first by describing the several methods they have used to deplete our capacity to exert self-control and then by describing their findings. Typical methods for limiting self-control have consisted of imposing time constraints and taxing working memory, either with digit-span tasks or the more demanding N-back task. Time constraints have elicited methodological criticism, as they cannot account of the effect of strength of preferences. Digit-span tasks have failed to tax cognitive capacity in any significant way, producing weak and conflicting results. N-back tasks have performed much better at that, and it is the one we have decided to use in our effort to contribute to the challenge of understanding the nature of altruism.

Our experiment applies an N-back task to an optimized 2AFC dictator game (DG). The optimization of the DG consists of choosing the amounts on each option in order to maximize the amount of information we can obtain about the underlying preferences of the participants. We started the development of this information efficient trial selection method for a previous experiment, and continued improving it for this one. By using this optimized DG task, we were able to i- analyze the effect of cognitive load on an individual level basis, and ii- look at the parameters of our participants' other-regarding preferences, instead of coarser measures such as total amount contributed. Like in our previous experiment, we evaluated decisions involving friends and strangers.

As expected, we found an incredible amount of individual variability. The effect of cognitive load on the parameters of other-regarding preferences was evenly distributed in all directions. Some became more prosocial, some less so. Some became more difference-averse, others less so. We confirm an interesting result from our previous experiment: more than 70% of participants show the same type of preferences towards friends and strangers, both under cognitive load and in the baseline condition. More than half (26 out of 47) show the same type of preferences in every scenario, both towards friends and strangers, and both with and without cognitive load. All of these individuals were prosocial. Additionally, almost half of the participants that showed a significant effect of cognitive load with both partners (16 out of 37) changed their preferences in the same direction when it came to friends as when it came to strangers. However, the direction of change was evenly distributed in all directions when we look at it between subjects.

A lot remains unanswered, and more experimentation needs to be carried out. What makes the effect of cognitive load go in one direction or another? Is it a stable effect or it depends on what someone is thinking and feeling on that day? We cannot say yet. One thing we can say, and that is that individual variability needs to be taken into account and for that, our information efficient trials can help.

Chapter 5

A Test of the Group Selection Theory of Altruism

5.1 Introduction

Altruism, especially altruism towards non-kin, has been a longstanding challenge to evolutionary biologists. Most acts of altruism entail a cost, and individuals that make sacrifices for others lose in relative fitness. This makes it less likely that their genes will pass on to the next generation. Lacking an opposing driving force to favor the genes that produce altruistic behavior, in time they will disappear from the population.

5.1.1 An Evolutionary Conundrum

Darwin recognized the challenge that altruism posed to his theory of natural selection. He wrote in The Descent of Man: “he who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature". His explanation of how this could be, reads: "a tribe including many members who . . . were always ready to give aid to each other and sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection" (Darwin, 1981).

This idea was the seed of the group selection theory of altruism. Group selection is one of many mechanisms by which any trait may develop. And altruism is but one emblematic, albeit contentious case. This theory has a long history that includes prominence, disrepute, and a veritable scientific controversy (Okasha, 2001). We will refer to this controversy in a future section. For now, Darwin's explanation simply suggests that, when groups of the same species compete for the same ecological niches, altruism may become an evolutionary stable trait. This could happen if groups with altruist members were more successful and reproduced faster than groups composed solely of selfish individuals. Then, group competition could, paradoxically, maintain a certain level of altruism in the global population.

5.1.2 History of a Controversy

Darwin's insight was intuitively very powerful, and by 1950s, it was mainstream in biology to interpret altruistic-like traits, such as the kamikaze bee sting, as adaptations for the benefit of the group, or the species.

The group selection idea, then most prominently represented by Wynne-Edwards [1962] and later by David Sloan Wilson [1977], started to fall in disrepute with two impactful critiques by George Williams [1966] and Maynard Smith [1964,1976]. They argued that group selection was i- biologically unlikely (it would only work with a very restrictive range of parameters), and ii- not necessary, as there were alternative explanations. Key among them, the kin selection theory.

5.1.2.1 Kin Selection

The most prominent alternative explanation to altruism was kin selection, proposed by Hamilton [1964a,b]. Hamilton noticed that animals that care for their offspring have a selective advantage over those who do not. Because the offspring carry the genes of the progenitors, genes that lead to altruistic behavior in progenitors may live on in the offspring. Caring for offspring is therefore an altruistic behavior that could have evolved by standard natural selection, through competition at the individual level.

Generalizing beyond offspring, kin selection posits that sacrificial behavior that benefits genetically related individuals might indeed be favored by natural selection. Hamilton thus expanded the concept of fitness to one that includes the fitness of relatives -to the extent that we share genetic information with them. If the beneficiary of a sacrificial act is a relative, they may carry the genes that produced the altruistic behavior. As mathematician and evolutionary biologist J.B.S Haldane reputedly joked: “I would gladly give up my life for two brothers or eight cousins” . Under the logic of kin selection, altruism is viable if there is a significant degree of genetic relatedness between the benefactor and the recipient.

Kin selection was widely seen as a superior explanation for altruism, despite the fact that it cannot explain altruism towards knowingly unrelated individuals, a behavior that is in fact prevalent in social species [de Waal, 1996, 2010, 2013].

5.1.2.2 Evolutionary Game Theory

Another alternative explanation, and one better suited to explain altruism towards non-kin, was ‘evolutionary game theory’, formalized with this name by Maynard Smith in 1982. These types of explanations model the fitness consequences of social interactions between pairs of individuals, but they require that the altruistic behavior ultimately benefits the benefactor, a condition that is not necessary either in kin or group selection.

The simplest case can be thought of as a Prisoner’s Dilemma. When an altruist interacts with a selfish individual, the later does well in detriment of the former. But when two altruists interact with each other, they do very well; and when two selfish individuals interact with each other, they do quite badly. Crucial to the evolution of altruism under this view is the frequency of interaction between the different types of individuals.

The pioneer of this line of thinking was Trivers, who proposed the influential concept of reciprocal altruism in 1971. He criticized Hamilton’s kin selection theory as not real altruism, as the genes that produce the supposed altruistic behavior perpetuate themselves, an idea that resonates with Richard Dawkin’s famous ‘Selfish Gene’. For him, real altruism requires that the parties are not genetically related. Crucially, it also requires that the benefactor and the beneficiary meet again and reverse roles, making the altruistic behavior actually profitable.

Evolutionary game theory was yet another alternative way of explaining altruism that did not rely on sacrificial acts for the good of the group but in good old self interest. By association, group selection fell in disrepute and was for the most part sidelined from mainstream evolutionary biology.

5.1.2.3 Group Selection

In defense of group selection, Sober and Wilson [1999] argue that the alternative explanations of altruistic behavior, such as kin selection and evolutionary game theory, are actually versions of group selection in disguise. They contend that the idea of group selection was discarded prematurely in the 1970s, and that Hamilton himself had come to share their view.

Under the logic of group selection, altruism is viable so long as there is a tendency for altruists to find themselves grouped together. By interacting with each other, altruists create a positive group effect that increases the fitness of everyone in the group, even the selfish members. Selfish members still have an individual advantage over altruist members within the group, but only by being in a group with several altruists can the entire group reproduce much faster, driving altruism to thrive in the global population.

In his book *Does Altruism Exist*, Wilson recounts an experiment conducted in his lab which provides a compelling example from the world of insects. It is a tale of water striders. Male water striders vary greatly in their aggressiveness towards females. Some act as rapists in human terms and attempt to mate with any female regardless of her receptivity. Others act as gentlemen, mating only when approached by females. In the water striders experiments, rapists outcompeted gentlemen for mates in mixed groups with both types of males. If within-group competition were the only evolutionary force in place, the gentlemen (‘altruism’) would ineluctably go extinct. However, because rapists (‘selfishness’) prevented females from feeding, they caused them to lay fewer eggs. The effect was so large that females in groups with only gentlemen laid over twice as many eggs as females in groups with all rapists. When water striders were allowed to move between groups, females ran away from rapists. Rapists were allowed to follow, but the net effect was a clustering of females around the gentlemen. Free movement thus creates enough variation among groups to maintain gentlemen in the population.

Crucially, non-random assortment of groups is the sine qua non for the existence of altruism under this view. Groups must differ in the proportion of altruists, or if migration is fluid between groups, altruists must tend to cluster together and interact more often. What precisely is the meaning of the word ‘group’ in ‘group-selection’ will be at the center of untangling this controversy.

5.1.2.4 Philosophy S.O.S. What is a Group?

In his paper "Why Won’t the Group Selection Controversy Go Away?" philosopher Samir Okasha argues that much of the controversy is due to conceptual confusion and provides much needed clarity to the debate.

Traditional group selection is presented as a higher-level analogy to individual natural selection. In this type of selection, individuals vary in a certain heritable trait that is relevant to fitness as they compete for survival. By natural selection, the traits associated with greater fitness are ‘selected’ because the individuals that carry those traits reproduce more. If group selection is the analogous process at the level of groups, then it should be groups that vary in a certain heritable trait, groups that compete, and groups that reproduce more or less, giving birth to new, distinct groups. This process describes how speciation works, but it applies as well to groups within a species.

Understood as a higher-level analogy to individual selection, traditional group selection presupposes groups that are reproductively isolated, spatially discrete, and last for several generations. These isolated, distinct groups reproduce giving way to new, distinct groups. But according to Sober and Wilson, this need not be. In their understanding, represented by Wilson's 'intra-demic' model of group selection (Wilson, 1977), groups need not be reproductively isolated, and may last for only a fraction of the lifespan of the group's members.

Maynard Smith fateful critique of Wilson's model centered on this very issue: there is no group reproduction, and therefore, no group heritability. It follows that there cannot be group selection. Sober and Wilson insisted that the effect of the group on its members' individual fitness is a form of group selection. According to them, in the case of altruism, there is group selection if i- there are groups that differ in fitness, and ii- there is a positive correlation between the fitness of the group and the proportion of altruists in it.

Okasha points out that the center of the disagreement is the concept of 'group fitness'. If group fitness means groups giving rise to new, distinct groups, then group heritability is absolutely necessary. If, on the other hand, group fitness means the average individual fitness of the group's members, then group heritability is not necessary, only individual heritability is. In that case, groups need not split into new, distinct groups; it is enough that their members reproduce faster.

So, is group selection really a higher order version of individual selection? Or is there a conceptual asymmetry? Do groups have to give birth to new groups for this theory to apply, or is it just the group's members that need to have a higher fitness?

According to Okasha, the idea of group selection as a higher-level analogy is terribly misleading. He notes that not even traditional group selection defined group fitness as groups giving birth to new groups, but as the average fitness of the individual group members. The two processes may very well be related, as groups with faster reproducing individuals may split into new groups more often than others. But he finds that group fitness understood as average individual fitness is sufficient for the theory to apply. He concludes that group heritability is irrelevant to group fitness and group selection. What matters is that the fitness of the individual cannot be entirely predicted by its own phenotype, but depends on properties of the group to which it belongs.

Okasha's philosophical analysis of the controversy may have been crucial in rehabilitating the idea of group selection in the evolutionary biology mainstream.

5.1.3 Tests of the Group Selection Theory

There has since been a resurgence of experimental testing of the group selection theory of altruism, particularly in humans. Different groups have tapped onto different economic games that pose a 'social dilemma' to the decision-maker, where cooperating brings a benefit to the group but is typically costly to the individual. Cooperation is not quite exactly altruism, but they are likely related.

The most commonly used game to study cooperation is the Public Goods game (PGG). The game consists of a group of people; each individual is given a monetary endowment and can make a voluntary contribution to a common pool. The common pool, often multiplied by a number greater than one by the experimenter to signify the goodness of the public good, is divided equally amongst the group members, regardless of their individual contributions. The classic parameters of the game are such that cooperation is costly. People still contribute positive amounts, a result that puzzled economists and biologists that held the classic view of presumed selfishness. For a review on PGG experimental results, see Ledyard's chapter in the Handbook of Experimental Economics, edited by Kagel and Roth (1995).

With a simple adaptation, the PGG is ideally suited to test the group selection theory of altruism. Pit groups against each other and have the competition be through voluntary contributions. Are groups with a higher proportion of altruists better prepared to compete against groups with mostly selfish people?

A hint to the answer came in 1993 from an Israeli orange groove. Erev et al, 1993 recruited male high school students to participate in an orange picking experiment with three different payment mechanisms: the individual, the collective, and group-competition reward conditions. The personal reward condition paid each group member according to what they had individually picked. The collective reward condition paid each member an equal part of the total picked. The group-competition condition divided the groups in dyads and gave a bonus to the dyad that picked more oranges. The collective payment resulted in a loss of 30% of production, but the group competition eliminated this loss in productivity all together. Interestingly, group competition was more effective at boosting productivity the more similar the groups were.

Back in the laboratory, Gunnthorsdottir and Rapoport, 2006 compare the PGG with and without group competition, considering two forms of payment: egalitarian and proportional. They confirm that group competition increases cooperation within the group, and that proportional payment yields the best results. Several other studies confirm these findings. They vary in the way they structure the competition between the groups. Some groups compete for a higher bonus (Burton-Chellew et al., 2010) or higher multiplier (Cárdenas and Mantilla, 2015). Some groups compete over a fixed prize (Abbink et al., 2012). In one case, only the hint of a group competition (pseudocompetition) generated this effect (Burton-Chellew and West, 2012). All of these studies point to the same conclusion: competition between groups increases cooperation within the group.

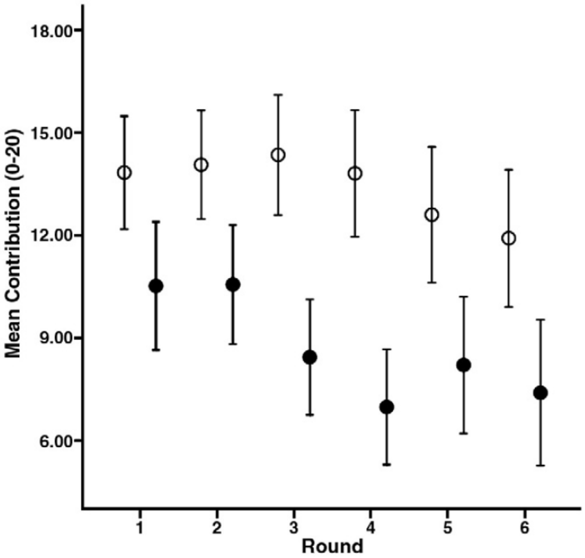

Figure 5.1 Effect of Group Competition on Intra-Group Cooperation, by Burton-Chellew. Reproduced from Burton-Chellew et al, 2010. Mean contributions in a 6-rounds Public Goods Game with (open circles) and without (black circles) group competition.

Other studies have cast doubt over the idea that cooperation in the PGG is due to prosocial preferences to begin with (Burton-Chellew and West, 2013; Kümmerli et al., 2010b). The PGG is not the game to measure prosocial preferences; rather, it is better suited to evaluate certain predictions about those preferences. Nonetheless, their results pose a serious challenge to established interpretations of group-competition PGG data. Kümmerli et al, 2010 note that, because it is impossible to make negative donations, any error in a PGG task will be expressed in the direction of pro-sociality. In costly cooperation, the equilibrium is zero contributions, and the data indeed shows that, after an initial period of naïve optimism, contributions see a steady decline towards a minimum – noteworthily, in the absence of group competition. If these minimal but positive contributions are better explained by the inability to reach such an extreme equilibrium, then a similar thing should happen when we take the game to the realm of profitable contribution. With the right multiplier, cooperating becomes profitable. The optimal strategy now shifts to full cooperation, even for selfish individuals. This means a contribution of 100% of the endowment. Under the theory of resistance to extreme strategies, we should see less than perfect donation under this condition. And this is indeed what they find. Contributions increased sharply but didn't reach 100%. In other words, they observed a generalized resistance to extreme strategies both under profitable and costly cooperation. In the case of profitable cooperation, they tested several conditions of 'induced cooperation'. Interestingly, most of those conditions of 'induced cooperation' consisted of some form of group competition.

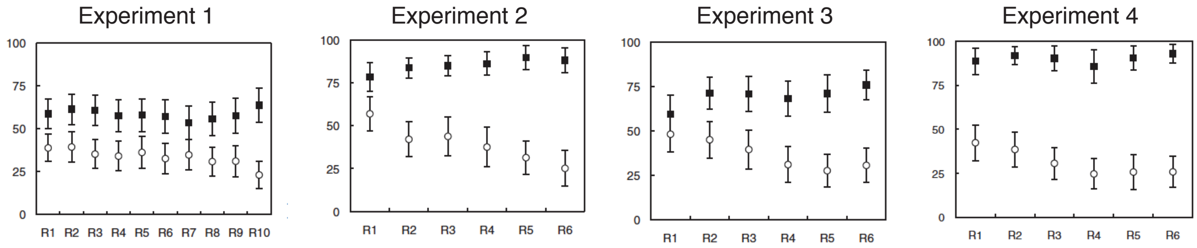

Figure 5.2 Effect of Group Competition on Intra-Group Cooperation, by Kummerli. Reproduced from Kummerli et al, 2010. Mean contributios under profitable cooperation (black squares) and costly cooperation (open circles). The 4 panels depict data from 4 different experiments that induce cooperation in different ways.

Burton-Chellew and West, 2013 dug an even deeper knife into the premise of social preferences underlying cooperation in these kinds of games. They test the classic PGG (without group competition) under three different information conditions: standard, reduced, or enhanced information. The standard condition informed participants, after each round, their personal gain and how much other members of the group had contributed. The reduced condition simulated a "black-box" situation, where participants were not even told that they were playing with others in a group. Lastly, the enhanced condition provided a detailed breakdown of how much each of the other group members had contributed, got back from the public good, and earned in that round. For both costly and profitable cooperation, detailed information about how others were personally benefitting from their generosity sunk cooperation.

5.1.4 The Need of Group Variability

In every case, there is a crucial aspect of the group selection theory that is overlooked: the fact that groups need to vary in their proportion of altruist members. Most of these studies use random assortment to form the competing groups. That is, the groups are formed randomly, and on each round of the PGG, group members are randomly shuffled. This is done so that people cannot track the reputations of other members of the group, especially at the moment of making 'altruistic' decisions. This design rules out the scope of reciprocity concerns, which have been proposed as a force that stabilizes cooperation (Nowak, 2006; Panchanathan and Boyd, 2004). The rounds then become independent, as if coming from entirely new groups each time. We contend that this practice undermines the very ability to test the theory. The theory of group selection requires groups that differ in a certain trait: the proportion of altruists. It is simply impossible to test using random groups that, by definition, do not vary in any significant way.

Moreover, the group selection theory of altruism does not state that competition between groups promotes altruistic behavior, although it might very well have that effect as well. Rather, it states that altruism persists in the population because groups with altruist members perform better than groups of selfish individuals. In the first scenario, altruism is understood as an adaptive strategy. Strategies might come to dominate over competing strategies through a dynamic process analogous to evolution. But that is a entirely different question. The question here is whether natural selection can produce a fixed trait, a fixed tendency to favor a certain strategy such as altruism, regardless of context, despite the fact that only in certain situations that trait is favored at the individual level.

Going back to the case of the water striders, the question is not whether rapists should become gentlemen when there is between-group competition. Or whether gentlemen should become rapists when females have nowhere to run. Rapists are rapists, and gentlemen are gentlemen. These are fixed traits. The question is not whether they change their behavior, but whether the fixed trait underlying that behavior remains in the population. The prediction of the group selection theory of altruism is not that competition will promote altruistic behavior but that groups of altruist members will outperform groups of selfish ones.

To test this hypothesis (not that group competition increases altruism but that altruist groups beat selfish groups) it is necessary to construct groups that vary in the proportion of altruist members, and have those groups -not random, equal groups- remain stable and compete. Stable groups provide at least a certain ecological validity when assessing competition between groups.

5.2 Methods

To test the hypothesis that “altruistic groups beat selfish groups” in humans, we designed a two-level structure of nested economic games (Fig.3). At the outer level, two groups –one altruistic and one selfish- competed in a Patent Race game. Inspired in the context of research and development, the Patent Race game consists of a winner-take-all mechanism where the side that invests more takes the entirety of the prize. At the inner level of the structure, each group participated in the Patent Race through a Public Goods Game (PGG). The investment that each group made in its attempt to win the prize was determined collectively via voluntary donations. At the ecological level, this design could describe the takeover of an encountered fruit tree by a group of moderately territorial primates.

In this design, it is crucial that groups vary in the percentage of altruist and selfish members. We achieved this via non-random group making. Before the main experimental task, we run a prescreening task, an optimized Dictator Game task, to measure each participant's altruistic inclinations. We then set up groups for the main experimental task with the aim of maximizing the difference in their average altruism. The groups then remained stable during the entirety of the main experimental task.

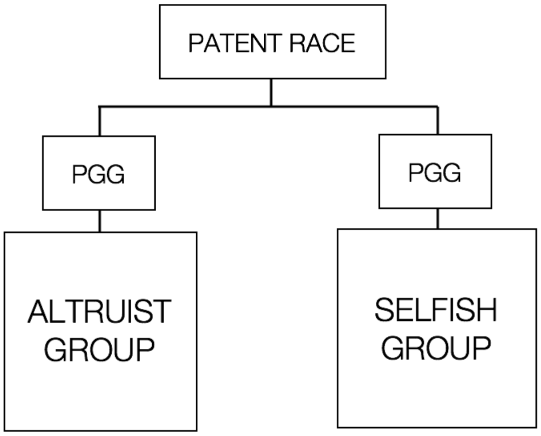

Figure 5.3 Patent Race through Public Goods Game Task. Two differentially altruistic groups compete for a prize (Patent Race) through voluntary contributions (PGG).

5.2.1 Subjects

Subjects were recruited through the Center for Experimental Social Science (CESS). Each person who signed up through the CESS system brought a friend along on the day of the experiment. The presence of the friend materialized ‘the Other’ in the prescreening dictator game (DG), which we used to obtain their altruistic inclinations. To maximize data collection, friends participated as decision-makers as well.

We ran this experiment in 10 sessions of 20 to 30 people approximately. Each session consisted of two parts. Everyone participated in the first part, the prescreening DG task, and almost everyone moved on to participate in the main experimental task, the Patent race through PGG, in the second part of the session.

5.2.2 Tasks

When participants arrived to the CESS Lab with their friends, they checked in with the experimenter and took a seat at one of the available computer stations. They waited until everybody arrived. Once everybody was in, the experimenter delivered instructions to the entire group, and answered questions. Participants signed their consent form.

5.2.2.1 Dictator Game

Prior to the main experimental task, participants completed an optimized dictator game, designed to measure their altruistic inclinations. We used the 1st generation DGM trials (See chapter 3 - A Systematic Characterization of Other-Regarding Preferences). Subjects completed 150 trials with the other being their friend, in 6 blocks of 25 trials each, in random order. To incentivize this task, one trial was randomly selected to enact payment. At the end of the DG task, the experimental code automatically fitted a Charness-Rabin social preference model and saved the estimated parameters into the subject' data file. As they finished this first part, they exited the lab, and waited in the hallway until the second part of the experiment could start.

During the waiting time, the altruism parameters of every subject were pulled out, averaged, and rank ordered. We then set up at least 4 groups of equal numbers of people, ranked from most to least altruistic, and dismissed the participants whose parameters didn't fit any group.

Two groups at a time were called back to participate in the second part of the experiment. The two groups at the extremes -the least and the most altruistic- played against each other, and the two groups in the center played against each other, in random order.

5.2.2.2 Patent Race through Public Goods Game

Altruistic and selfish groups proceeded to play 10 rounds of the patent race through PGG. On each round, each group member received 10 monetary units (MU). They could contribute any amount they wanted to the common cause of winning the prize and keep the rest for themselves. The prize consisted of the total sum of the contributions by both groups, multiplied by 4. As per the patent race rules, the group that contributed the most in a specific round, won the prize. The prize in that round would be divided equally among the members of the winning group. To incentivize the task, one round would be randomly selected and enacted for payment. Total earnings consisted of their share of the prize, if their group won, and the MU they kept for themselves. Finally, each MU was converted to dollars at a rate of 1/4. This task was coded and presented using the z-tree software (Fischbacher, 2007).

5.2.3 Variables of Interest

If the group selection theory of altruism is true, we expected to see that altruist groups beat the selfish group at the patent race more often than not. We considered the possibility that during the 20 rounds of the game, selfish groups could manage to increase cooperation to be able to beat altruist groups at the patent race. This would not be a problem. The presence of fixed traits need not preclude the emergence of adapted strategies that may operate on top of them. The question is whether, regardless of circumstantial strategies adopted by the groups, groups with the fixed trait in question consistently outcompete groups without it.

If selfish groups adapt their strategy and increase cooperation in order to compete with the altruist group, there are two possible scenarios. Groups may converge in behavior and become indistinguishable, in which case the data would argue against the need for group competition to maintain altruism in the population, and perhaps even question the need to conceptualize altruism as a fixed trait. Alternatively, groups may adapt their strategy but not converge, and altruist groups may manage to maintain an advantage over selfish groups. These results would provide strong evidence in favor of selective advantages of altruist groups, and thus, of the existence of altruism in the population

5.2.4 Incentive Compatibility

Both tasks – the DG and the Patent Race through PGG- were paid according to the subjects' choices to assure incentive compatibility. For the DG task there was also a payment accrued due to the friend's choices. DG trials consisted of amounts for self and other that reached \$100. And to this we add the show-up fee of \$15. To make the experiment economically feasible, we took a series of measures that maintained incentive compatibility: First, either the friend's choices or the subject's (both involve payments for both people) were selected for payment, randomly. Second, DG amounts remained as is with a 10% chance, and otherwise would be converted to dollars at a rate of 1/8. For the patent race task, because the prize could potentially be so big but started with smaller numbers, we used a conversion rate to dollars of 1/4.

After the fact, we noticed however that there was a mismatch between the range of payment of each task. If they kept the entirety of the endowment in the patent race task, they received a maximum of 2,5 dollars, which ex-post might have been a bad choice of design. In comparison to the payment from the DG task, this is negligible. This might drive selfish individuals to make higher contributions, not because that is a trade-off that they would normally do, but because the alternative being so little, it is not worth considering. This design error should make it more difficult to see the hypothesized effect.

5.3 Results

5.3.1 Subjects

252 people participated in this experiment in 10 sessions of typically 20 to 30 people each. Outside this range, one session had 16 people, one 32, and another one 36. All of the 252 participants completed the prescreening dictator game task. Of them, 228 moved on to participate in the main experimental task, the Patent Race through PGG. Each session created typically 4 groups, with their size raging from 4 to 6 members each. One session with 36 people in 6 groups resulted in the loss of control of the experimental conditions. The session was completed despite the disruption, and immediately flagged as problematic and considered for exclusion. Our final data set consists therefore of a total of 192 people who participated in the patent race through PGG. All the analyses, however, were run again including the problematic session. Results were for the most part the same, except that they were noisier. Since they did not add any illuminating elements, we do not include them here.

5.3.2 Individuals and Groups Visualization

We fitted a Charness-Rabin social preferences model to 252 participants who completed the DG task. The experimental code, written in Matlab, automatically fitted the model using a genetic algorithm (ga function). The two parameters of interest -weight for other when better off (ρ) and weight for other when worse off (σ) - were bounded to the range -1 to 1, where the values are expected to lie. The groups for the main experimental task were arranged using these parameters. In Figure 5.4(a) we show a later fit without bounds on the parameters using matlab's function fmincon. As a result of unbounding the fitting process, 18 participants out the 252 had new estimated parameters for ρ and/or σ outside the range of what seems reasonable for the Charness-Rabin model. Of them, 15 participated in the following patent race task. Because it is unclear what the true preferences of these subjects are, they add to the noise and hinder the emergence of the hypothesized effect. We excluded these 18 subjects from Figure 5.4(a) for visualization purposes.

Of these 252 people, 228 formed groups and participated in a group competition match. There were 2 matches in 9 of the 10 sessions, and 3 matches in the discarded session. This amounts to 18 matches, and the potential addition of 3 more. In Figure 5.4(b), we show an example match.

5.3.3 Group Competition

On each session, 4 groups were set up by ranking and grouping participants according to their altruistic preferences. The most and least altruistic groups competed against each other, leaving the other two more moderate groups to compete amongst themselves. The pairs played a patent race through PGG for 10 rounds.

In terms of round by round data, both groups started with their lowest contribution in the first round, increased sharply by the second round, and maintained an elevated contribution throughout the final round. Interestingly, the more altruistic groups on average made consistently higher contributions in every single round, but the difference is small. A 2-way mixed-effects ANOVA, with rounds as a within-subjects factor and group as a between-subjects factor, revealed a significant main effect of group ($F(1)$=11.45, p=0.006), a significant main effect of rounds ($F(9)$= 7.81, p=0.003), and no interaction

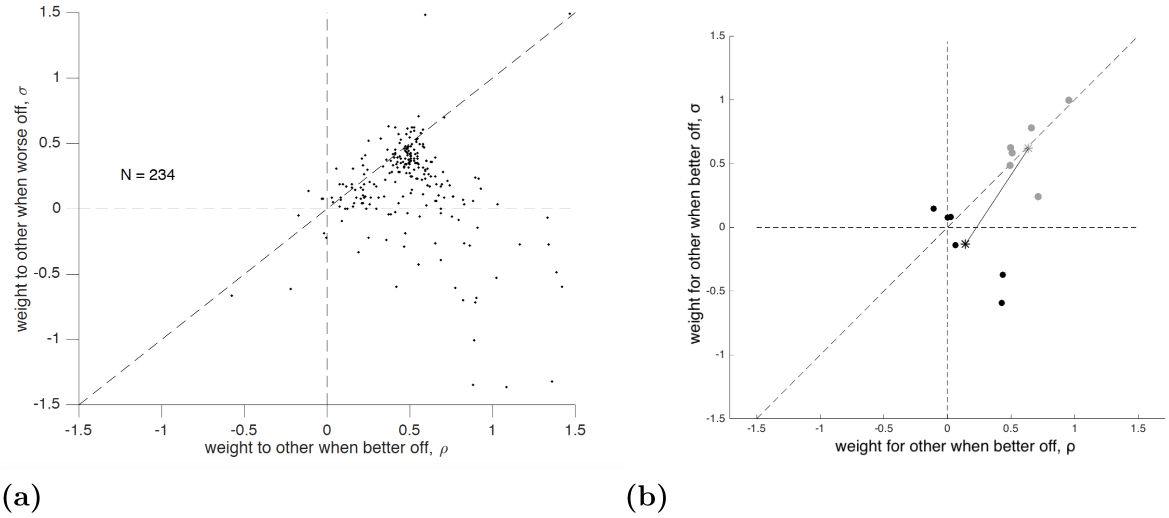

Figure 5.4 Individual Preferences and Example Group Competition. **(a)** Individual parameter estimates of 234 of the 252 subjects that completed the DG prescreening task. **(b)** Example group competition match, with individual group members in filled circles and the group average in asterisks. Grey denotes the more altruistic group, black denotes the least altruistic group. The text describes the percentage of rounds won by the more altruistic group (% won), the difference in average altruism (D Alt), and t-test results for whether we can reject that the groups are equal in their altruism (H) with it's textitp-value.

(F(9)=0.67,p=0.73).

We then ran two series of post-hoc t-test, one without and one with the Holm-Bonferroni correction for family-wise error. This is because Holm-Bonferroni assumes that all tests are independent, and this is most likely not the case, producing an over-correction. Without correction, in the first 5 rounds the difference is not statistically significant, but in the last 5 rounds it becomes significant and remains so until the end of the game (ps<0.035). With the Holm-Bonferroni correction, most rounds were not significant, only the 8th round showed a significant difference (p=0.0012<alpha-HB=0.005).

On average, across all rounds, the more altruistic group contributed 8.72 ± 0.2 MU whereas the least altruistic group contributed 8.05 ± 0.4 MU. This difference was not statistically significant ($t(34)$= 1.49, p= 0.15). See Figure 5.6.

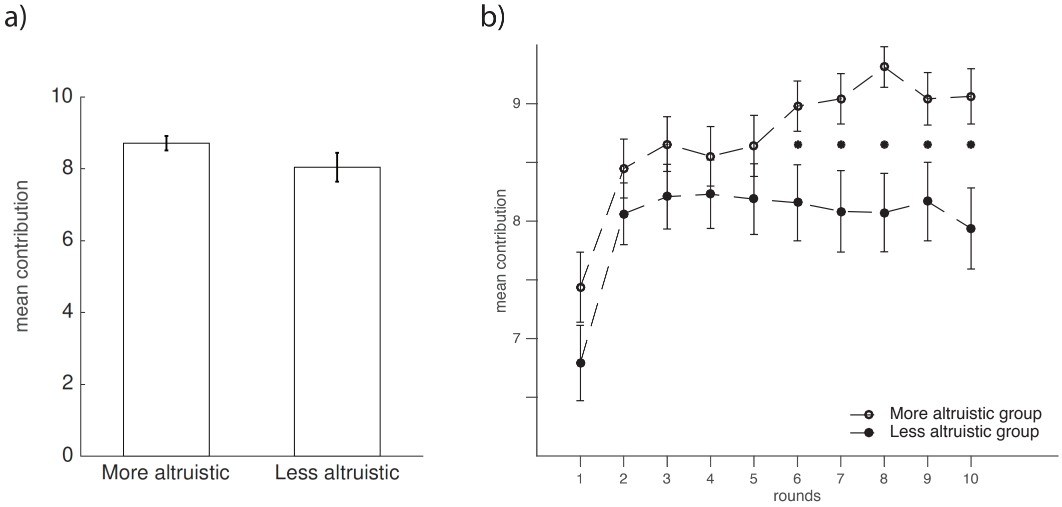

Figure 5.5 Group Contributions. **(a)** Mean cross-rounds group contributions, with s.e.m. **(b)** Mean round-by-round contributions by group. Open circles represent the more altruistic groups, and black circles represent the least altruistic groups. Asterisks mark rounds where the groups differed as per uncorrected t-tests. Only the 8th rounds was different with Holm-Bonferroni correction.

This parsimonious picture at the aggregate level hides a much more variable reality at the individual level of each match between competing groups. Of the 18 matches, in 50% of the cases the more altruistic group did not win most rounds. It did win more than loose, and when it won, it tended to do so by a larger margin, but the pattern failed to show statistical significance on several tests.

The more altruistic group won in 50% of the matches, 15% were ties, and 35% of the time, the least altruistic group won the most rounds. In the 50% of matches where the more altruistic group won, they did so with an average of 83% of rounds as victories. In the 35% of matches where the least altruistic group won, they did so with an average of 63% of rounds as victories. Overall, the average percentage of rounds won by the more altruistic group, including ties, was 63%. While suggestive, the hypothesis that the more altruistic groups win most of the time (>50%) is not significant on a Sign test, p=0.607.

A Spearman rank-order correlation coefficient between the difference in group altruism and the percentage of rounds won by the most altruistic group was not significant either, $p = 0.52$. This implies that we fail to see a monotonic rank-order relationship between the variables. A non-monotonic relationship between them is not necessarily discarded. See Figure 5.6.

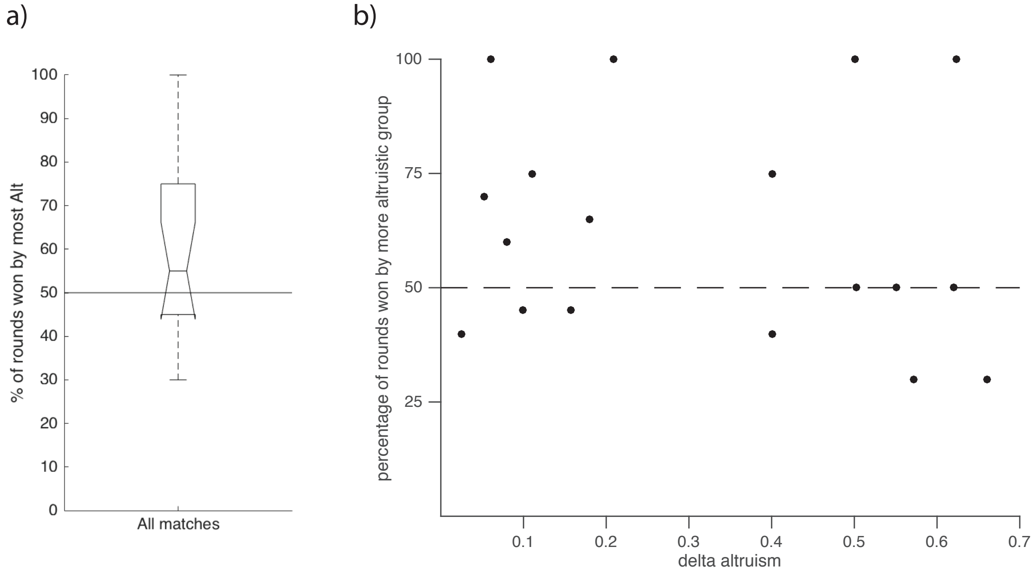

Figure 5.6 Percentage of Altruistic Victories. **(a)** Interquartile range of the percentage of rounds won by the more altruistic group. **(b)** Scatter plot of the differnece in altruism between the two competing groups and the percentage of rounds won by the more altruistic one. If more altruistic groups perform better than selfish ones, the percentage of rounds won by the former should be above 50%.

These results are suggestive but not convincing. There are several reasons why this could be. The first reason is a miscalculation of the incentive structure of the experiment as a whole, resulting in a mismatch between the first and second tasks. Specifically, if subjects did not contribute at all in the patent race through PGG task, they would take home only $2.5 for that task. But the prescreening DG task was designed to ensure that everyone would take home at least $20, possibly more, and up to $100 in a small percentage of cases. By making the reward for non-cooperation almost neglectable compared to the total payment, these experimental parameters excessively incentivized cooperation, even and especially amongst those participants who might have been inclined to choose the non-cooperative option. In other words, the “selfish” group was inadvertently incentivized to cooperate, by making the selfish reward excessively small.

Another reason for the unconvincing results is that, by having ran both tasks in the same session, we could not optimally tailor groups. In some sessions, the “selfish” group was not selfish at all, it was just less altruistic. Furthermore, in some cases, the difference in altruism between groups was very small or minimal. This could be overcome if all participants completed the prescreening task first and were selectively asked to come back for a second session, with groups formed considering all available individuals. This would have allowed us to form groups with actually selfish individuals instead of simply less altruistic ones.

Other errors such as fitting a bounded model with the ga function instead of an unbounded model using the fmincon function contributed to adding noise to the data. All of these things undermined the emergence of the hypothesized pattern.

5.4 Conclusion

In this chapter I explored a controverted evolutionary hypothesis for the existence of altruism towards strangers, namely, the group selection theory of altruism. I reviewed this and alternative hypotheses, such as kin selection and evolutionary game theory, and recounted the scientific discussion between the proponents of each theory. Kin selection is the most well-known hypothesis, but it cannot explain altruism towards genetically unrelated individuals, let alone complete strangers. Evolutionary game theory overcomes this limitation but requires altruism to be ultimately profitable. Group selection is free of these caveats but has not managed to recover from emphatic detraction and garner widespread adherence.

Initial opposition to group selection was a reaction against an overreliance on explanations of certain behaviors as being done for the “benefit of the group”. In the context of that shift, it was attacked as a gross and outdated misunderstanding of the theory of natural selection. Its defenders argued that it was cast away too quickly, its detractors themselves having misunderstood it. I presented philosopher Samir Okasha’s intervention in detangling the controversy. He concludes that group selection defenders “arrive at the correct position by faulty reasoning” and points to the key concept underlying the confusion: what constitutes a ‘group’ for the purposes of group selection. Is group selection a higher order version of individual selection, with ‘group’ being the unit of selection? Or is there a conceptual asymmetry? He concludes the latter.

I then reviewed experimental research that tested the validity of the group selection theory in humans. The results are inconclusive for several reasons. The most important one being an error as to what the theory predicts. To the best of my knowledge, all of the previous studies focus on whether group competition increases within-group cooperation. Focused on that, the groups are formed randomly by the experimenters, and as a result, they are on average the same. They typically find that group competition increases cooperation, but this is not what the theory actually says. The theory says that more altruistic groups will outperform selfish ones, and for that, it is paramount to form groups that differ in their level of altruism. This is my main contribution.

I then presented the experimental design and results. I leveraged on some of our previous work aimed at measuring altruism with high precision at the individual level. With that tool at my disposal, I was able to form differentially altruistic groups. My groups remained stable during the experiment, and this can be argued to be problematic as well. In terms of group formation, there are two separate instances to have in mind. One is the initial random group formation, and the second is random reshuffling of group members on each round. Random group formation is a problem, random reshuffling is not. Ideally, I should have formed groups non-randomly (one with more altruistic members and the other with selfish ones) and then reshuffled group members on each round maintaining the differential characteristic of each group. That precludes reputation building coming into play.

Despite correcting the conceptual error of using equal groups, my results were not convincing either. They seemed to suggest the validity of the theory, with more altruistic group performing slightly better, but too many statistical tests failed to be significant. This could be due to errors on our part, which were plenty. The gravest was a mismatch between task compensations that may have overly incentivized cooperation amongst noncooperators, erasing the potential difference between the two groups. The overall conclusion is that more research needs to be done with better thought out incentives, but crucially, keeping the differential group design.

Some suggestions for further research include systematically varying the incentive structure, such as the cost and benefit of cooperation. Do altruist groups perform better under certain conditions but not others? One of the most powerful critiques to the group selection theory of altruism is that the parameters under which altruism would prevail are so extreme that it is unlikely the driving force behind it. Under which parameters would we see that altruist groups start winning? Another variable that could be explored is group composition. For this, I would have to split the experiment in 2 sessions, and conduct all the prescreening sessions first, and once everybody's altruism parameters are obtained, carefully tailored groups could be called back to participate in the patent race through PGG. By running both tasks within the same session, I was limited to form groups with the people available in that particular session. In some of those sessions, there were very few selfish individuals to make actually selfish groups. I had to content myself with just less altruistic ones. This might have eroded the emergence of the hypothesized result. Lastly, I could incorporate random reshuffling on each round, maintaining differential grouping.

Chapter 6

Conclusion

In this thesis, I explored a series of questions about the nature of altruism. The motivation was simple: if we want to strive for the best possible organization of society, we need to know what we are working with. Excessively idealistic conceptions will lead to impossible demands and ultimately backfire. In hoping for the best for our fellow humans, we cannot deny who we are, with all of our virtues and our limitations. What is, after all, to love ideal humans and deprecate real ones?

Any committed approach must start with an honest investigation of the variables at play. Our dispositions towards others are at the center of the answer. And so, I set out to investigate what our other-regarding preferences look like. What started as a simple exercise in model comparison branched out an unexpected methodological detour when I was confronted with the question of what type of trials to use. I noticed that while there was enormous progress being made in terms of measurements (such as the introduction of linear and stepwise budget sets,) there were still limitations and ad hoc choices made by the experimenters. These ad hoc choices may be good choices, but I decided to pursue a principled method, nonetheless.

I described this methodological adventure in the first chapter, before any of the experimental ones. I decided to present it separately in order not to distract from the experimental results, despite the fact that the development of this trial selection method evolved in parallel to the experimental projects and constituted a central aspect to them. For instance, I would not have been able to aim at an individual level analysis of the data had it not been for these information maximizing trials.

6.1 The Methodological Contribution

There are several aspects that can be discussed in relation to this methodological work. The first is that, of course, the quality of the results provided by the selected trials are as good as the elements that went into the trial selection method in the first place. Those elements are mainly two: the candidate trials themselves, and the candidate models that the trials are aiming to distinguish.

The candidate models were selected based on wide popular use, or conceptual simplicity, or symbolic meaning. I considered the Charness-Rabin model, a Cobb-Douglas, and a Rawlsian model. The Charness-Rabin includes the possibility of a Fehr-Schmidt model. Both are widely used. The Cobb-Douglas is conceptually simple but improves on the linearity of the previous two models by sake of its convexity. The Rawlsian model contributes meaning to the list, as it describes the radical egalitarian. See chapter 2 - An Information Efficient Trial Selection Method, section 2.3.2 - Candidate Models of Other-Regarding Preferences for the mathematical form, and chapter 3 - A Systematic Characterization of Other-Regarding Preferences, section 3.2.5 - Models of Other-Regarding Preferences for their meaning.

Other models would have been excellent candidates but were left out. In the future, we should consider them. One such model is the lexicographic, which describes a realistic selfish person: someone who would never make sacrifices for others but would not mind benefitting them if it were free. Another such model is Cox's Egocentric Altruism, a constant elasticity of substitution (CES) function modified to allow for different weights for others depending on whether the decision maker is better or worse off. CES functions have the limitation that the weights for others must be positive, but the Cobb-Douglas model shares this trait and was included anyhow. Many times, the CES function is reduced, depending on some of its parameters, to a linear, a Cobb-Douglas, or a Rawlsian function, and for this reason, I decided against including it.

In any case, a subpar list of models will produce subpar results. And while I believe our list of candidate models mostly covers the type of utility functions proposed for non-strategic interactions, I feel that I have failed to make justice especially to the lexicographic model. And so, other lists of models, particularly including this one, could greatly contribute to our understanding of other-regarding preferences.

In regard to the candidate trials, while we worked with an incredibly long list of them, our method suffered from the same limitation that it purported to solve: a certain ad hoc element to it. The selection of trials was principled, but the list of candidate trials was not. Therefore, this method could never claim to have found the 'ultimate trials.' Other, better candidate trials may be tested. This may seem like an obsessive point but note that when we compared the performance of 75 first-generation trials with 80 ad hoc Cendri trials (Hutcherson et al., 2015), the latter did significantly better. It can be argued that the comparison was not fair because the first-generation trials were aimed at distinguishing a uniform population, but were evaluated on an experimentally informed one, and what they had gained in identifying experimentally irrelevant agents, they had lost in identifying the relevant ones. But the Cendri trials had no cudgel at all, no sophisticated methods to back them, even if flawed. They were chosen by invoking no less than the wisdom of the naked eye, the natural intuition that we have and lets us know what extracts information and what does not. OK, calm down, the second-generation trials did perform better – once we incorporated experimental priors and the assumption of overlap of mutual information of different trials… So what are we make of this? Well, I would not now start advocating for the method of the naked eye, but other lists of candidate trials, perhaps using a little bit of ancient wisdom, could make an improvement.

Beyond the list of candidate models and candidate trials, a more interesting improvement could come from bridging different types of tasks. There is a tension between finding a model that describes a certain type of preferences very well and a model that describes all preferences well. Those with the latter inclination would like to find the model that describes behavior from dictator games, plus ultimatum games, plus prisoner dilemma games, plus trust games (Blanco et al., 2011). The logic behind this is that these are all types of social decisions. But as we have mentioned in the Social Preferences section of the Introduction chapter, there might be all sorts of variables coming into play differentially in some games more than in others (such as 'saving face'), and not at all in yet other games. After all, we do not lump together social preferences, plus risk preferences, plus intertemporal preferences, etc. Why not? The same logic of the desirability of an overarching theory applies. My inclination has been to be guided by the psychological process that I aim to study. And so, when it comes to altruism, I have stuck to the dictator game. This, however, does not mean that an overarching model could not be credibly proposed in the future, and when that happens, I would like to see some type of DGM trials give it the best chance.

Lastly, it is worth mentioning a kindred spirited endeavor by Breitmoser, (Breitmoser, 2013). Breitmoser also set out to measure social preferences with high precision but took a different approach. He noted several categories of models according to where the noise is assumed to be coming from. In the family of structural models, he listed: *random behavior* models, which add noise at the moment of the response; *random taste* models, which introduce noise in the preference parameters; and *random utility* models, where the utilities of all options are perturbed randomly before a perfect maximizer picks the best among the perturbed options. Breitmoser also evaluated regression models, which are free of primitives such as preferences and maximization. He argues that getting this aspect right is crucial in correctly estimating social preferences. He evaluated different models by analyzing precision (descriptive adequacy) and robustness (predictive adequacy). He found that structural models avoid overfitting, and that random utility models are the most robust. The salient difference between Breitmoser's aim and mine is that he centers on the choice aspect of the decision, and that is valid, while I center of the preferences themselves. It would be interesting to explore the possibility of bringing the two approaches together.

6.2 The Experimental Contributions

With the description of the novel trial selection method behind, I turned to present the experimental projects. The three projects addressed the following three guiding questions: 1) How altruistic are we? 2) Where does it come from? Is it self-control? And 3) If we are altruistic, how is it possible, from an evolutionary point of view?

The first experimental project tackling how altruistic we are consisted of a systematic characterization of other-regarding preferences. I considered a shortlist of models for model comparison first, and once I had a clear winner, I focalized on parametrization. An overwhelming majority of my first sample followed the Charness-Rabin model. When I focalized on parametrization, I found that the vast majority of them had prosocial preferences. This means that they had positive weights for others both when better and worse off, albeit less so when worse off. Only a small minority followed the Fehr-Schmidt model of difference aversion, with positive weights for other when better off and negative weights when worse off. This finding was particularly interesting because of the symbolic meaning of the difference averse preferences and because even when researchers fit Charness-Rabin models to their data, they still tend to call it ‘difference aversion’. But status-dependent preferences do not imply difference aversion, as the prosocial instantiation of Charness-Rabin model can attest. Difference aversion requires antisocial dispositions towards others when worse off. The prosocial Charness-Rabin preference is sensitive to being better or worse off, but values efficiency more, and would not pay to see others hindered.

Another interesting finding in this project was that, for the most part, subjects showed the same type of preferences towards friends and strangers. Those who were competitive towards strangers, were so towards friends as well. Those who were prosocial towards friends, were so towards strangers too. The same can be said of difference-averse subjects. Of course, the magnitude of the parameters was different with friends and strangers. As a whole, participants tended to be more altruistic towards friends. However, the type of preference, the overall pattern of behavior, remained the same.

In future research, it would be interesting to see if this pattern holds if we test a bigger sample of people or include the lexicographic model. The lexicographic model is a strong competitor to Charness-Rabin when it comes to strangers, but not when it comes to friends. The current trials may not be suited to discover a better performing lexicographic model. Future runs of the trial selection method should definitely include it. If it did describe choices towards strangers better, the observation that people have the same type of preferences towards friends and strangers may very well fall. For now, the Charness-Rabin model has done a superb job at describing other-regarding preferences.

Another interesting direction where this research could develop is in testing ideological framings. I use the word ‘ideological’ to distinguish framing effects informed by broad cosmovisions about how the world should work, or how humans should behave. The 2AFC version of the dictator game rid us from undesirable framing effects that arise in coarser versions of the game, where there is a fixed amount of money and the subjects are asked how much of their endowment they wished to ‘give’ to their partner. As we have mentioned in the Conclusion section of chapter 3 - A Systematic Characterization of Other-Regarding Preferences, the mere including of the option to ‘take’ changes the results dramatically, moving the modal choice to not offering anything at all (Bardsley, 2008; List, 2007). This alternative framing including the option to take is illuminating but does not provide a solution to the problem of undesirable framings at all. The 2AFC DG overcomes this problem, but that does not mean that other, desirable framings are worth studying. What would happen if, before the DG task, participants read about the supposed virtues of socialism or capitalism? What if the framing where implicit, and they read a story about someone being selfless versus another story about someone being selfish? What if both cases had versions where having that trait led to success versus failure?

In general terms, it is still unclear what determines our altruistic inclinations. Is it driven primarily by a sense of empathy? How many motives actually feed into giving behavior? It has been proposed that several motives, such as warm glow or guilt - in addition to altruism, modulate giving (Bolle et al., 2012). Other competing motives may come into play. How does envy, for instance, affects giving behavior? And what is relationship of all of that with pure altruism? Bolle et al, 2012 stand apart in their attempt to separate motives. All too often, researchers are eager to test interesting phenomena with complicated designs that end up mixing all sorts of variables together. And even more often still, they jump to conclude or think that they are seeing or using difference averse preferences, when it reality, they are not (Binmore and Shaked, 2010).

The second experimental project dealing with whether altruism is the product of self-control consisted of a cognitive load manipulation. It has been a long-standing question where altruism in humans comes from. Historically, the assumption was that socialization was required to bring altruism about. In the same way that manners needed to be taught and inculcated, selflessness needed to be too. Self-control of selfish impulses was a necessity. But more recent theories suggest that a certain level of altruism may be in our very nature. From biological theories of when and how altruism is favored by natural selection to recent experimental results from cognitive psychology experiments, the suggestion is that altruism may be as much of an impulse as selfishness was previously thought to be.

Upon reviewing the literature and the different methods for implementing cognitive load, I chose to use the N-back task in simultaneous to the basic social decision-making task. With an easy and a hard condition, I set out to find out if altruism increased or decreased under cognitive load. The view that humans are selfish and require self-control to engage in altruism predicts that under cognitive load, we will revert to our selfish ways. The opposite view predicts that altruism will be unleashed. I leveraged on the power that the information maximizing trials provided to consider the possibility that I would find a great deal of individual variability.

And indeed, that is what I saw. I first confirmed previous findings: most people were prosocial, and most people had the same type of preferences towards friends and strangers. But the effect of cognitive load went in all directions! Interestingly, of the people that showed a significant effect of cognitive load both when it involved friends as well as when it involved strangers, half had that effect go in the same direction with both partners. And lastly, of the subjects that showed the same type of preference in every condition, towards friends as well as stranger, under cognitive load as well as on the baseline condition, all of them were prosocial.

This dispels simplistic theories about the nature of altruism. Not only because the effect of cognitive load showed great individual variability, but also because people did not radically transform from being altruistic to being selfish, or vice versa. Future research is required to understand what determines the direction of change of the effect of cognitive load on social preferences. Another interesting research avenue is what is the relationship of this effect to theories involving social norms (Cohen et al., 2001; Hawkins et al., 2019)? And if norms have become intimately ingrained in our nature through a process of gene-culture coevolution (Axelrod, 1986; Gintis, 2011; Richerson et al., 2010), what does that imply for altruism as an impulse as opposed to it being a product of self-control?

Finally, I presented the third and last experimental project. This project tested the group selection theory of altruism, a controverted evolutionary hypothesis for the existence of altruism towards strangers. This theory poses that groups with altruist members perform better than groups with selfish members when there is competition between them. In its original formulation, performance is measured in reproductive terms. I adapted it to a modern human realm, leaving reproduction aside and turning to economic performance. Presumably, economic performance can serve as a proxy for reproductive success before the advent of the contraceptive pill and modern medicine, whereby child mortality greatly decreased.

In this chapter, I introduced alternative explanations for altruism, such as ‘kin selection’ and ‘reciprocal altruism’, and the limitations that each have at accounting for pure altruism towards strangers. The group selection theory was dominant before these other two theories became widely accepted and favored. The group selection theory, associated to a bad interpretation of it (the idea that individuals engage in behavior for the benefit of the group), fell in disrepute. I then recounted the debate amongst its defenders and detractors, and the intervention of a philosopher that brought much needed clarification to the discussion. The group selection theory’s reputation began a process of rehabilitation, the success of which remains to be seen.

I then reviewed recent experiments that tested the group selection theory, transferred to the realm of economic games with humans. They showed promising results. Group competition increased cooperation within the groups. However, there was a severe flaw in their design: they all had equal groups. While increased cooperation within groups when there is between-group competition is suggestive, the theory absolutely requires that groups vary in their proportion of altruist members. I was able to set up differentially altruistic groups by leveraging on the power that our trial selection method provided in measuring altruistic preferences at the individual level.

I erred, however, in the precise design of the incentives structure, pushing non-cooperators to cooperate and look more similar to cooperators. This diluted the potential emergence of the hypothesized effect. Nonetheless, a subtle effect emerged in the predicted direction, with altruistic groups cooperating more in the end, but by a small margin. Altruistic groups won more often than lost, but the results are not overwhelmingly convincing. More research is called for to come to a definite conclusion.

There are several improvements to our design that come to mind. One of course is correcting the incentive structure. Another is varying the conformation of the groups. Is more altruistic always better? Or is there such a thing as too much altruism? For that, the prescreening of the individual altruistic preferences of all of the participants should be completed before the groups are conformed and the first group competition session begins. As things stand right now, my results do not allow me to discard or corroborate the validity of the group selection theory of altruism.

All in all, this thesis only dents at some of the aspects of altruism. Many more aspects remain to be studied. Hopefully, in the future many others will join me in this endeavor. For now, I can say that the selfishness imperative may be too hard. Animal expressions of altruism may be key in dispelling this myth. As it stands, the field of ethology is providing convincing evidence that we do not need socialization to be good natured. Famed primatologist Frans De Waal dedicated an entire book to finding examples of altruism in the animal world. He focused primarily on primates but ventured into other types of animals as well (De Waal, 1996)). Now we can breathe a sigh of relief. Because all those cute videos that we saw on social media of animals risking life and limb to save another, be it of the same species or not, or behaving in ways that we thought only humans could, have an academic explanation that can put our minds at rest. Not to draw too much of a parallel, but studies in children may also contribute to break the myth of the selfish imperative (Barragan et al., 2020; Burkart et al., 2007; Warneken et al., 2007; Warneken and Tomasello, 2006, 2009). As Binmore and Shaked put it: "it is uncontroversial that people care about others to some extent" (Binmore and Shaked, 2010).

References

Abbink, K., Brandts, J., Herrmann, B., and Orzen, H. (2012). Parochial altruism in intergroup conflicts. *Economics Letters*, 117(1):45–48. ISBN: 0165-1765 Publisher: Elsevier.

Acerbi, L. and Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In *Advances in neural information processing systems*, pages 1836–1846.

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77. ISBN: 0020-6598 Publisher: JSTOR.

Afriat, S. N. (1972). Efficiency estimation of production functions. *International economic review*, pages 568–598. ISBN: 0020-6598 Publisher: JSTOR.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723. ISBN: 0018-9286 Publisher: Ieee.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546. ISBN: 0012-9682 Publisher: JSTOR.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477. ISBN: 0013-0133 Publisher: JSTOR.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753. ISBN: 0012-9682 Publisher: Wiley Online Library.

Axelrod, R. (1986). An evolutionary approach to norms. *The American political science review*, pages 1095–1111.

Baker, C. L., Goodman, N. D., and Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the thirtieth annual conference of the cognitive science society*, pages 1447–1452.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133. ISBN: 1386-4157 Publisher: Springer.

Barkow, J. H., Cosmides, L., and Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press, USA.

Barragan, R. C., Brooks, R., and Meltzoff, A. N. (2020). Altruistic food sharing behavior by human infants after a hunger manipulation. *Scientific reports*, 10(1):1–9.

Benjamin, D. J., Brown, S. A., and Shapiro, J. M. (2013). Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, 11(6):1231–1255. ISBN: 1542-4766 Publisher: Oxford University Press.

Berger, P. L. and Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Penguin Uk. Issue: 10.

Binmore, K. and Shaked, A. (2010). Experimental economics: Where next? *Journal of economic behavior & organization*, 73(1):87–100.

Bisley, J. W. and Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual review of neuroscience*, 33:1–21.

Blanco, M., Engelmann, D, and Normann, H.T. (2011). A within-subjects analysis of other-regarding preferences. *Games and Economic behavior*, 72:321–338.

Bolle, F., Breitmoser, Y., Heimel, J., and Vogel, C. (2012). Multiple motives of pro-social behavior: evidence from the solidarity game. *Theory and decision*, 72(3):303–321.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193. ISBN: 0002-8282.

Braams, B. R., Güroğlu, B., de Water, E., Meuwese, R., Koolschijn, P. C., Peper, J. S., and Crone, E. A. (2014). Reward-related neural responses are dependent on the beneficiary. *Social Cognitive and Affective Neuroscience*, 9(7):1030–1037. ISBN: 1749-5024 Publisher: Oxford University Press.

Breitmoser, Y. (2013). Estimation of social preferences in generalized dictator games. *Economics Letters*, 121(2):192–197.

Brizendine, L. (2006). *The female brain.* Broadway Books.

Burkart, J. M., Fehr, E., Efferson, C., and van Schaik, C. P. (2007). Other-regarding preferences in a non-human primate: Common marmosets provision food altruistically. *Proceedings of the National Academy of Sciences*, 104(50):19762–19766.

Burton-Chellew, M. N., Ross-Gillespie, A., and West, S. A. (2010). Cooperation in humans: competition between groups and proximate emotions. *Evolution and Human behavior*, 31(2):104–108. ISBN: 1090-5138 Publisher: Elsevier.

Burton-Chellew, M. N. and West, S. A. (2012). Pseudocompetition among groups increases human cooperation in a public-goods game. *Animal Behaviour*, 84(4):947–952. ISBN: 0003-3472 Publisher: Elsevier.

Burton-Chellew, M. N. and West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, 110(1):216–221. ISBN: 0027-8424 Publisher: National Acad Sciences.

Buss, D. (2015). *Evolutionary psychology: The new science of the mind.* Psychology Press.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and brain sciences*, 12(1):1–14. ISBN: 1469-1825 Publisher: Cambridge University Press.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

Cappelletti, D., Güth, W., and Ploner, M. (2011). Being of two minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology*, 32(6):940–950. ISBN: 0167-4870 Publisher: Elsevier.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869. ISBN: 0033-5533 Publisher: Oxford University Press.

Cohen, M. D., Riolo, R. L., and Axelrod, R. (2001). The role of social structure in the maintenance of cooperative regimes. *Rationality and Society*, 13(1):5–32.

Cornelissen, G., Dewitte, S., and Warlop, L. (2011). Are social value orientations expressed automatically? decision making in the dictator game. *Personality and Social Psychology Bulletin*, 37(8):1080–1090. ISBN: 0146-1672 Publisher: Sage Publications Sage CA: Los Angeles, CA.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

Cox, J. C. and Sadiraj, V. (2005). Direct tests of models of social preferences and introduction of a new model. Technical report, University of Arizona working paper.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, 47(2):448–74. ISBN: 0022-0515.

Cárdenas, J. C. and Mantilla, C. (2015). Between-group competition, intra-group cooperation and relative performance. *Frontiers in behavioral neuroscience*, 9:33. ISBN: 1662-5153 Publisher: Frontiers.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80. ISBN: 0938-2259 Publisher: Springer.

Darwin, C. (1872). The expression of emotions in animals and man. *London: Murray*, page 11.

Darwin, C. (1981). The descent of man, and selection in relation to sex. 1871. *Princeton: Princeton UP*.

Dawkins, R. (1989). *The selfish gene.* Oxford university press.

De Waal, F. (1996). *Good natured. The origins of right and wrong in humans and other animals.* Harvard University Press.

De Waal, F. (2010). *The age of empathy: Nature's lessons for a kinder society.* Broadway Books.

De Waal, F. and Waal, F. B. (2013). *The bonobo and the atheist: In search of humanism among the primates.* WW Norton & Company.

Ekman, P. E. and Davidson, R. J. (1994). *The nature of emotion: Fundamental questions.* Oxford University Press.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4):583–610. ISBN: 1386-4157 Publisher: Springer.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869. ISBN: 0002-8282.

Erev, I., Bornstein, G., and Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, 29(6):463–478. ISBN: 0022-1031 Publisher: Elsevier.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic inquiry*, 41(1):20–26. ISBN: 0095-2583 Publisher: Wiley Online Library.

Fareri, D. S., Niznikiewicz, M. A., Lee, V. K., and Delgado, M. R. (2012). Social network modulation of reward-related signals. *Journal of Neuroscience*, 32(26):9045–9052. ISBN: 0270-6474 Publisher: Soc Neuroscience.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868. ISBN: 1531-4650 Publisher: MIT Press.

Feingold, A. (1994). Gender differences in personality: a meta-analysis. *Psychological bulletin*, 116(3):429. ISBN: 1939-1455 Publisher: American Psychological Association.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178. ISBN: 1386-4157 Publisher: Springer.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5):1858–1876. ISBN: 0002-8282.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.

Fox, R. (1989). *The search for society: Quest for a biosocial science and morality*. Rutgers University Press.

Gintis, H. (2011). Gene–culture coevolution and the nature of human sociality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1566):878–888.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., and Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research*, 21(2):241–253. ISBN: 0885-7466 Publisher: Springer.

Glimcher, P. W. (2011). *Foundations of neuroeconomic analysis*. Oxford University Press USA.

Glimcher, P. W. (2014). Value-based decision making. In *Neuroeconomics*, pages 373–391. Elsevier.

Glimcher, P. W., Dorris, M. C., and Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and economic behavior*, 52(2):213. Publisher: NIH Public Access.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, 30.

Gunnthorsdottir, A. and Rapoport, A. (2006). Embedding social dilemmas in intergroup competition reduces free-riding. *Organizational Behavior and Human Decision Processes*, 101(2):184–199. ISBN: 0749-5978 Publisher: Elsevier.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4):367–388. ISBN: 0167-2681 Publisher: Elsevier.

Hacking, I. and Hacking, J. (1999). *The social construction of what?* Harvard university press.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16. ISBN: 0022-5193 Publisher: Elsevier.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1):17–52. ISBN: 0022-5193 Publisher: Elsevier.

Hauge, K. E., Brekke, K. A., Johansson, L.-O., Johansson-Stenman, O., and Svedsäter, H. (2016). Keeping others in our mind or in our heart? Distribution games under cognitive load. *Experimental Economics*, 19(3):562–576. ISBN: 1386-4157 Publisher: Springer.

Hawkins, R. X., Goodman, N. D., and Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in cognitive sciences*, 23(2):158–169.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78. ISBN: 0002-8282.

Hobbes, T. (1651). *Leviathan.* New York: Oxford University Press.

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, 17(66):159–174. ISBN: 0013-0427 Publisher: JSTOR.

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, 87(2):451–462. ISBN: 0896-6273 Publisher: Elsevier.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795. ISBN: 0162-1459 Publisher: Taylor & Francis.

Krajbich, I., Bartling, B., Hare, T., and Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature communications*, 6(1):1–9. ISBN: 2041-1723 Publisher: Nature Publishing Group.

Kümmerli, R., Burton-Chellew, M. N., Ross-Gillespie, A., and West, S. A. (2010a). Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences*, 107(22):10125–10130. ISBN: 0027-8424 Publisher: National Acad Sciences.

Kümmerli, R., Burton-Chellew, M. N., Ross-Gillespie, A., and West, S. A. (2010b). Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences*, 107(22):10125–10130. ISBN: 0027-8424 Publisher: National Acad Sciences.

LeDoux, J. E. and Damasio, A. R. (2013). Emotions and feelings. *Principles of neural science*, pages 1079–1094. Publisher: McGraw-Hill New York.

Ledyard, J. O., Kagel, J. H., and Roth, A. E. (1995). Handbook of experimental economics. *Public Goods: A Survey of Experimental Research*, pages 111–194. Publisher: Princeton University Press NJ.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of economic dynamics*, 1(3):593–622. ISBN: 1094-2025 Publisher: Elsevier.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3):482–493. ISBN: 0022-3808 Publisher: The University of Chicago Press.

LoBue, V., Nishida, T., Chiong, C., DeLoache, J. S., and Haidt, J. (2011). When getting something good is bad: Even three-year-olds react to inequality. *Social Development*, 20(1):154–170. ISBN: 0961-205X Publisher: Wiley Online Library.

Locke, J. (1690). *An essay concerning human understanding*, volume 1668. London.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American economic review*, 90(2):426–432. ISBN: 0002-8282.

Louie, K., Khaw, M. W., and Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, 110(15):6139–6144. ISBN: 0027-8424 Publisher: National Acad Sciences.

Morgenstern, O. and Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., and Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, 75(1):73–79. ISBN: 0896-6273 Publisher: Elsevier.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

Oberholzer-Gee, F. and Eichenberger, R. (2008). Fairness in extended dictator game experiments. *The BE Journal of Economic Analysis & Policy*, 8(1). Publisher: De Gruyter.

Okasha, S. (2001). Why won't the group selection controversy go away? *The British journal for the philosophy of science*, 52(1):25–50. ISBN: 1464-3537 Publisher: Oxford University Press.

Panchanathan, K. and Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432(7016):499–502. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Pareto, V. (1906). *Manual of Political Economy. 1971 translation of 1927 edition*. New York: Augustus M. Kelley.

Pinker, S. (2003). *The blank slate: The modern denial of human nature*. Penguin.

Piovesan, M. and Wengström, E. (2009). Fast or fair? A study of response times. *Economics Letters*, 105(2):193–196. ISBN: 0165-1765 Publisher: Elsevier.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 10(2):59–63. ISBN: 1364-6613 Publisher: Elsevier.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302. ISBN: 0002-8282 Publisher: JSTOR.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2013). Rand et al. reply. *Nature*, 498(7452):E2–E3. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., and Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature communications*, 5(1):1–12. ISBN: 2041-1723 Publisher: Nature Publishing Group.

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Richerson, P. J., Boyd, R., and Henrich, J. (2010). Gene-culture coevolution in the age of genomics. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8985–8992.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71. ISBN: 0013-0427 Publisher: JSTOR.

Schulz, J. F., Fischbacher, U., Thöni, C., and Utikal, V. (2014). Affect and fairness: Dictator games under cognitive load. *Journal of Economic Psychology*, 41:77–87. ISBN: 0167-4870 Publisher: Elsevier.

Segerstråle, U. C. O. (2000). *Defenders of the truth: The battle for science in the sociology debate and beyond*. Oxford University Press.

Sell, A., Tooby, J., and Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, 106(35):15073–15078. ISBN: 0027-8424 Publisher: National Acad Sciences.

Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of consumer Research*, 26(3):278–292. ISBN: 1537-5277 Publisher: The University of Chicago Press.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439(7075):466–469. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Smith, A. (1759). *The theory of moral sentiments*. Penguin.

Smith, J. M. (1964). Group selection and kin selection. *Nature*, 201(4924):1145–1147. ISBN: 0028-0836 Publisher: Springer.

Smith, J. M. (1976). Group Selection. *Quarterly Review of Biology*, 51(2):277–283.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge university press.

Sober, E. and Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press. Issue: 218.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017. ISBN: 1053-8119 Publisher: Elsevier.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., and Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, 112(5):1619–1624. ISBN: 0027-8424 Publisher: National Acad Sciences.

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, 6(5):363–375.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Västfjäll, D., Kirchler, M., and Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498(7452):E1–E2. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Tooby, J. and Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. ISBN: 1593856504 Publisher: Guilford Press.

Tricomi, E., Rangel, A., Camerer, C. F., and O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, 463(7284):1089–1091. ISBN: 1476-4687 Publisher: Nature Publishing Group.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1):35–57. ISBN: 0033-5770 Publisher: Stony Brook Foundation, Inc.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., and Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, pages 1874–1882.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., and Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biol*, 5(7):e184.

Warneken, F. and Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *science*, 311(5765):1301–1303.

Warneken, F. and Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in cognitive sciences*, 13(9):397–402.

Watson, J. B. (1924). *Behaviorism.* New Brunswick: Transaction.

Waytz, A., Zaki, J., and Mitchell, J. P. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *Journal of Neuroscience*, 32(22):7646–7650. ISBN: 0270-6474 Publisher: Soc Neuroscience.

Williams, G. C. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought Princeton University Press.* Princeton University Press.

Wilson, D. S. (1977). Structured demes and the evolution of group-advantageous traits. *The American Naturalist*, 111(977):157–185. ISBN: 0003-0147 Publisher: University of Chicago Press.

Wilson, D. S. (2015). *Does altruism exist?: culture, genes, and the welfare of others.* Yale University Press.

Wilson, E. O. (1975). *Sociobiology: The new synthesis.* Harvard University Press.

Wynne-Edwards, V. C. (1962). Animal dispersion: in relation to social behaviour. Technical report.