KATHOLIEKE UNIVERSITEIT
LEUVEN

FACULTY OF SCIENCE
Department of Biology
Ecology, evolution and biodiversity conservation

# A study of prosocial punishment in humans using experimental games

by

Loren PAUWELS

Promotor: Prof. T. Wenseleers
Biology Department, KULeuven

Co-promotor: Prof. S. Dewitte
Economy Department, KULeuven

Dissertation presented in fulfillment of the requirements for the degree of Master of Science in Biology

Academic year 2011-2012

# Acknowledgements

Ik voel me enorm bevoorrecht dat mijn omgeving bestaat uit een aantal heel bijzondere mensen, die me elk op hun manier hebben bijgestaan gedurende mijn studentenjaren. Dankzij hen kan ik iets doen wat ik graag doe.

In de context van dit eindwerk zou ik de volgende personen in het bijzonder mijn dank willen betuigen:

Mijn promotor, Prof. Tom Wenseleers, en mijn co-promotor, Prof. Siegfried Dewitte. Zij hebben me de mogelijkheid gegeven om een onderwerp te bestuderen binnen het veld van mijn interesse, ondanks het feit dat een dergelijke thesis misschien een iets minder voor de hand liggende optie was voor een biologiestudent. Eveneens zijn zij degenen die me bijstonden met begeleiding. Hun kritische feedback, stimulerende ideeën en discussies hebben me erg vooruit geholpen. Dank u wel, het was een erg boeiende ervaring.

Een woord van dank is zeker ook op zijn plaats aan Aiste, Nathalie en Anouk, die me wegwijs hebben gemaakt in Authorware en steeds bereid waren om me te helpen wanneer ik op problemen botste bij het programmeren.

Ik zou ook graag iedereen van het labo van socioecologie en sociale evolutie willen bedanken voor de thesis- en niet thesisgerelateerde goede raad en (behoorlijk talrijke) momenten van ontspanning. Ik heb van mijn tijd op het lab genoten, al hou ik er wel een koffieverslaving en een wantrouwen voor jacuzzi's aan over.

Niet op z'n minst een hele grote dankjewel aan mijn ouders en familie om me zowel financiëel als mentaal te blijven steunen en me de kans te geven om mijn weg te vinden. Mama, papa, Joske en opa, jullie aanmoedigingen hebben me vaak frisse moed gegeven. Samen met mijn (behoorlijk geweldige!) vrienden zijn jullie degenen waar ik niet zonder had gekund.

Ale, de uitwerking die ge op mij hebt is ondoorgrondelijk. Door u heb ik een betere versie van mijzelf kunnen zijn de voorbije jaren (jammer maar helaas, normaalgezien ben ik nog erger dan dit). Danku om het met mij uit te houden. Ik hoop dat ge het nog een beetje langer volhoudt, zodat ik de kans krijg om ook zoveel voor u te betekenen. Ik hou van u.

# Contents

I. Introduction ..... 1

1 Approaches in the study of human behaviour ..... 1

1.1 Evolutionary psychology ..... 1

1.2 Human behavioural ecology ..... 2

1.3 Game theory ..... 3

1.3.1 Classical game theory ..... 3

1.3.2 Evolutionary game theory ..... 3

1.4 Dual inheritance theory ..... 4

2 Cooperation and prosocial behaviour in human societies ..... 5

3 Mechanisms for promoting prosocial behaviour ..... 6

3.1 Kin and group selection ..... 6

3.2 Direct reciprocity ..... 8

3.3 Indirect reciprocity ..... 9

3.4 Costly signalling ..... 9

3.5 Punishment ..... 10

4 Experimental and behavioural economics ..... 11

4.1 Experimental game theory ..... 11

4.2 Examples of games ..... 11

4.2.1 Ultimatum game ..... 12

4.2.2 Prisoner’s dilemma ..... 12

4.2.3 Public goods game ..... 13

5 Altruistic punishment ..... 14

5.1 Strong reciprocity and morality ..... 16

5.2 Proximate mechanisms ..... 16
5.3 Ultimate explanations ..... 17
5.3.1 Inequity averse utility functions ..... 17
5.3.2 Cultural group selection ..... 18
5.3.3 Hitchhiking of altruistic norms ..... 19
5.4 Opposition to prevailing explanations of altruistic punishment ..... 19
5.4.1 Misfiring hypothesis ..... 20
5.4.1.1 The artificiality of one-shot, anonymous encounters ..... 20
5.4.1.2 Misfiring in other animals ..... 21
5.4.2 Criticism of cultural group selection models ..... 22
5.4.3 Selfish punishment ..... 24
5.4.4 Theoretical models on selfish punishment ..... 24
6 Aims of the thesis and hypotheses ..... 25
II. Materials and methods ..... 27
1 Experimental setup ..... 27
1.1 Game course ..... 28
1.2 Programming and data analysis ..... 29
III. Results ..... 33
1 Descriptive analysis of punishment behaviour ..... 33
2 Descriptive analysis of individual contribution ..... 34
3 Multiple regression analysis of individual contribution ..... 35
4 Multiple logistic regression analysis of punishment behaviour ..... 37
5 Multiple linear regression analysis of individual payoffs ..... 47
6 Multiple linear regression analysis of the group’s total payoff ..... 50

IV. Discussion ..... 52
1 The effect of punishment on cooperation levels ..... 52
2 The effects of the cost of punishment ..... 53
2.1 Frequency of punishment ..... 53
2.2 Profile of the punisher ..... 54
3 Observed punishment strategies ..... 54
3.1 Cooperators as punishers ..... 54
3.2 Loners as punishers ..... 55
3.3 Defectors as punishers ..... 55
4 The relative success of different strategies ..... 56
5 The effect of punishment on the success of the group ..... 57
6 Concluding remarks ..... 58
V. Summary ..... 59
VI. Samenvatting ..... 60
VII. References ..... 61
VIII. Addendum ..... 68

# I. Introduction

Human cooperation is a truly puzzling subject that has managed to both fascinate and confuse researchers from various backgrounds. Punishment has been put forward as one of the explanations for man's high levels of cooperation; however this behaviour itself seems to raise even more questions. The first sections (1-3) of this introduction will allow the reader to get acquainted with the study of human behaviour and human cooperation. Next, an experimental framework commonly used to study human cooperation is introduced (section 4). Finally, we further expand on punishment and its proposed explanations (section 5). In section 6, we specify the aims of our study.

## 1 Approaches in the study of human behaviour

Human behaviour is currently attracting a lot of interest in diverse fields, ranging from biology, psychology, economics, philosophy, anthropology, and several others. Consequently, different approaches to the study of human behaviour have emerged, each offering a different framework for interpretation. Two of those approaches, evolutionary psychology and human behavioural ecology, try to explain human behaviour from an adaptationist point of view, but give a different emphasis and have some theoretical and methodological differences. A third approach we will discuss is game theory, which is used as a conceptual and mathematical framework to make predictions about optimal behavioural outcomes by economists as well as biologists. To conclude this brief overview, dual inheritance theory is introduced. This theory emphasises the importance of cultural processes and suggests that they alter the process of selection.

### *1.1 Evolutionary psychology*

Evolutionary psychology is a relatively recent term, but the foundations of evolutionary psychology (and behavioural ecology, for that matter) have been laid out centuries ago with the introduction of the Darwinian way of thinking (Daly and Wilson, 2004). Evolutionary psychology is usually solely concerned with the study of *Homo* sapiens and focuses on the psychological mechanisms that evolved for specific situations, as well as on the evolutionary context in which they evolved (Pinker *et al.,* 1992; Daly and Wilson, 2004). According to evolutionary psychologists, the outcome of evolution is a modular brain where each module evolved to deal with a certain fitness problem. Also, when identifying the relevant selective pressures involved in the evolution of the brain modules, these researchers

think in terms of the ancestral evolutionary environment (AEE) (Bowlby, 1969; Smith *et al.,* 2000). By consequence, they argue that most of our modern-day society environments differ so radically from the AEE that 'misfiring' of our brain should be very common. This means that a lot of our behaviour will probably be maladaptive in our current environment because of the evolutionary time lag; our brain's modules are optimized for the ancient or Pleistocene environment, but not for the current one. This provides an explanation for the fact that we learn to fear spiders and snakes more readily than guns, even though guns are a much bigger source of danger in our modern society: Our psychological machinery for fear is stuck in the past in which poisonous snakes and spiders posed a substantial threat (Öhman and Mineka, 2001). A downside of this approach is that evolutionary psychologists hypothesize about brain processes, but can only test this by looking at behavioural outcomes. However, often these behaviours are thought to be the result of a complex combination of modules and their corresponding brain processes and drawing any inferences will be difficult unless there is a one to one mapping. In addition, it is not possible to get to know exactly all of those selective pressures that were acting in the past, which means we can never be certain about what behaviours would really have been adaptive in the AEE (Foley, 1995).

### *1.2 Human behavioural ecology*

In human behavioural ecology, behavioural outcomes are investigated instead of psychological mechanisms. In contrast with evolutionary psychology, behavioural ecology measures the current adaptiveness of a behavioural trait by studying individual differences in reproductive success. Behavioural ecology focuses on explaining behaviour as a function of ecological and social context (Smith *et al.,* 2001). For humans, the same models as for other animals are used to discover the factors that play a role for a certain behaviour's fitness. A behavioural ecologist assumes a fitness optimisation model, which is another point of contention between them and evolutionary psychologists, who instead assume an optimal adaptation to an AEE but not to the modern world. Real biological data often tends to confirm these optimization models. For example, optimisation models succeed to predict how a mother's workload would influence the time between two births (Jones and Sibly, 1978). Humans have evolved many ways of social learning and knowledge transition and human behavioural ecologists make the assumption that humans can behave very flexible in response to new environments (Smith *et al.,* 2000).

## *1.3 Game theory*

The foundations of game theory were laid out by Von Neumann and Morgenstern (1944), offering a mathematical tool to predict the outcomes of strategic interactions (Fudenberg and Tirole, 1991; Osborne and Rubinstein, 1994; Camerer, 2004). Very rapidly, this tool was applied to the field of economics (I, 1.3.1) and later on, it also became a powerful method in biology (I, 1.3.2).

### 1.3.1 Classical game theory

Game theory studies situations in which people have to make decisions of which the consequences, also formulated as the payoff or the outcome, depend on decisions made by others. This situation can be presented as a game, where all persons involved in the game are players and where the decisions that players need to make can be seen as strategic choices. The desirability of the outcome of their behavioural strategy is referred to as utility, which is calculated as a utility function (Camerer, 2004). In the standard theory, this utility function corresponds to the expected payoff (Page *et al.,* 2000). Since all decisions in the game are made simultaneously, information on fellow-players' strategic choices is lacking and players will have to apply iterated reasoning. An assumption made in classical economical game theory is that humans act rationally and are aware of each others' rationality: all players will have certain beliefs regarding the strategies that other players will adopt and they themselves will choose a strategy so that, given those beliefs, their expected utility is maximized (Crawford, 1997). John Nash (1950) predicted that the game would have a stable point at which rational players would no longer adjust their strategy, because any change in strategic play would give rise to a lower utility (Nash Jr, 1950). This point is called the Nash equilibrium and is a key concept in game theory (Camerer, 2004).

### 1.3.2 Evolutionary game theory

Actually, this game analogy is applicable to a wider range of situations. Other than persons, the players could for example be genes, firms or nations, whereas strategies could be genetically coded instincts, bidding behaviours, legal strategies or wartime battle plans. Outcomes represent anything that players value, for example mating opportunities, power, food, prestige, money or territory (Camerer and Fehr, 2004). Inspired by classical game theory, Maynard Smith adapted the theory to analogous 'game situations' in nature where different behavioural strategies exist in a population and the success of those strategies depend on the

frequency of other strategies (Smith, 1982; Smith, 1986). A first important difference with classical game theory is that instead of utility, fitness is maximized. Also, natural selection is the process that drives this optimization, not rationality. Consequently, an equilibrium situation does not require any assumption about human rationality. In fact, Maynard Smith argues that since it is hard to decide on what is rational, we cannot suppose people to act rationally. With the Hawk-dove game as a classic example, evolutionary game theory was born. This theory can be used to analyse cases in which the fitness of a phenotype is dependent on the frequency of its own and other phenotypes in the population. Also, Smith introduced the concept of an "evolutionary stable strategy" (ESS), a concept analogous to the Nash equilibrium. One can state that a strategy is evolutionary stable if a population displaying a certain phenotype or behavioural strategy cannot be invaded by a rare mutant adopting a different strategy (Smith and Price, 1973).

### *1.4 Dual inheritance theory*

Boyd and Richardson's (2009) statement: "Something makes our species different, … that something is cultural adaptation" seeks to affirm a common sense notion that man distinguishes itself from all other animals. To investigate the validity of this notion, it is crucial to find out in what ways culture works to establish this presumed difference. Importantly, rather than suggesting a false dichotomy between nature and culture, it must be emphasised that culture itself is an adaptive product of genetic evolution, enabling humans to acquire adaptive traits trough social learning (Boyd and Richerson, 1988). Culture can be defined as those aspects of "thought, speech, behaviour and artefacts", which can be socially learned and transmitted (Cavalli-Sforza and Feldman, 1981). However, once culture has been established through social learning mechanisms and the appropriate psychological machinery, as argued by dual inheritance theorists, it becomes a potent evolutionary force itself. Cultural evolution can alter the selective environment and interact with genetic evolution, leading to a process referred to as gene-culture coevolution (Henrich and Henrich, 2007). For example, a study on the post-marriage residential cultural practices of the Sino-Tibetan-speaking hill tribes of Thailand unveils that the genetic diversity of mitochondrial DNA (which is passed on trough the mother) is much greater in the patrilocal villages than in the matrilocal villages, providing evidence that cultural evolution can influence genetic evolution (Oota *et al.,* 2001).

Social learning gives rise to a new system of inheritance of cultural traits (also called 'memes' by R. Dawkins (1983)), which are passed with a certain amount of error,

just like genes are. Some argue, though, that cultural evolution is different from genetic evolution because the possible modes of transmission are more diverse, ranging from purely vertical (from parents to offspring), to oblique (from members of the parental generation other than the parents) or horizontal (between individuals of the same generation), whereas genes are usually only passed on vertically. Departing from this line of thought, new models were developed to include cultural transmission. Among others, this was done by Richerson and Boyd (2004), arguing that social learning can speed up the evolutionary process, rendering group selection scenarios plausible (Richerson and Boyd, 2004; Boyd and Richerson, 2009; Boyd *et al.,* 2011). In section I, 3.1, we offer a brief introduction to kin and group selection theory and in section I, 5.3.2, we will discuss cultural group selection models further.

## 2 Cooperation and prosocial behaviour in human societies

Human societies can appear to be rather special in comparison to other animal societies. Humans communicate through spoken language, go to concerts for diversion, practice religion, live by certain moral standards and function in complex and highly organized communities. Also, humans are often described as 'hyper-sociable' (Boyd and Richerson, 2009). We support and care for the weak members of society, engage in trades, have large scale conflicts in which individual soldiers can sacrifice their lives, establish systems for norm enforcement and offer help to unrelated strangers. Unlike in other social species, cooperation can sometimes thrive under circumstances where the established evolutionary explanations would not seem to suffice (section In 5.1).

A cooperator is defined as someone who bestows a fitness benefit to one or more individuals at his own dispense. A defector is an individual who does not cooperate (or cooperates less than his fair share), but is potentially able to gain the benefit of others cooperating. West and Griffin (2007a) describe cooperation as "a behaviour which provides a benefit to another individual (recipient), and which is selected for because of its beneficial effect on the recipient", to emphasize that cooperative behavior includes all altruistic and some (but not all) mutually beneficial behaviours.

Cooperation has since long been an evolutionary puzzle and specifically, the above-mentioned patterns of human cooperation constantly create new question marks. If one takes Darwin's original evolutionary theory as a starting point, where an individual maximizes its own fitness by maximizing its survival and reproduction, it

seems difficult to explain why individuals would engage in any altruistic behaviour. In other words, naively one would predict Darwinian adaptation to always lead to selfish individuals (Nowak, 2006). As we will see, however, this is not quite the case, as many theories and mechanisms have been proposed that can promote cooperation and prosocial behaviour.

## 3 Mechanisms for promoting prosocial behaviour

In the following paragraphs, we will present some of the most influential theories which have been proposed to promote prosocial behaviour. In order to explain cooperation and other behavioural traits, evolutionary biologists are guided by one ground rule: If a form of behaviour is adaptive, then it must, directly or indirectly, bestow a fitness benefit to the actor (Fig. 1). Below, in section I, 3.1, we will discuss theories based on indirect fitness benefits, whereas in sections I. 3.2, I. 3.3 and I. 3.4 we will review various theories based on direct fitness benefits.

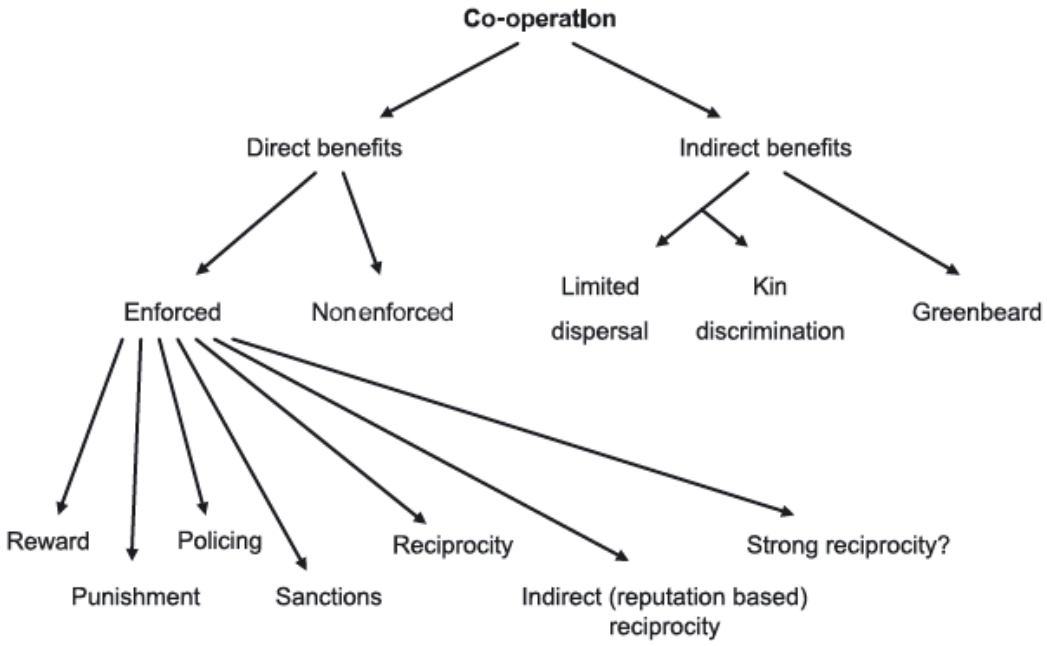



**Figure 1: A classification for the explanations of cooperation as proposed by West Griffin et al. (2007a), by categorizing the mechanisms according to whether they provide direct or indirect fitness benefits.**

### *3.1 Kin and Group selection*

Inclusive fitness theory, also known as kin selection theory (Maynard Smith, 1964), highlighted the fact that that the indirect benefits of increasing the fitness of strongly

related individuals could possibly outweigh the direct costs of reducing one's own fitness (Haldane, 1955). This idea led to a revision of the Darwninian concept of fitness, which had traditionally always been formulated in terms of personal reproduction or survival. W. D. Hamilton redefined it as inclusive fitness, which is the sum of direct and indirect fitness, and where indirect fitness gains can be obtained by helping related individuals, who carry copies of the actor's own genes (Hamilton, 1964). Hamilton also introduced what is now known as Hamilton's rule, "$r.b_r > c_a$", which expresses the net condition under which an altruistic gene would be able to spread in the population. In this rule, r is the degree of relatedness, $b_r$ is the reproductive benefit to the recipient of the behaviour and $c_a$ is the reproductive cost to the actor that displays the (altruistic) behaviour. Hamilton's rule states that the relatedness must exceed the cost-to-benefit ratio of the altruistic act (r>c/b) for the altruistic behaviour to evolve.

Like kin selection theory, group selection emanates from the idea that selection operates on more than just the level of the individual. Selection can occur on the level of genes (kin selection), but also on higher-order units such as social groups, species and multispecies communities (Wilson, 1997). Traditionally, group selection is thought to help the evolution of cooperation trough differential extinction between groups (Wynne-Edwards, 1963). For instance, a group of selfish individuals would go extinct faster because they would overexploit their environment, or in a conflict between rivalling groups, the group that masters cooperative warfare would be the strongest competitor. This theory often led to the general misconception that behaviours would develop for the good of the group, population or even species. Also, there are several problems with this kind of group selection, such as the requirement that dispersion is extremely low or non-existent and the fact that altruist groups would be very susceptible to social parasitism. In the late seventies, however, a number of group selection theoreticians emerged who insisted that group selection should not written off (Price, 1972; Wilson, 1975; Colwell, 1981; West *et al.,* 2007a). They redefined the concept of a group as a (temporary) within-population association of individuals and emphasised that selection was active on multiple, not necessarily equally important, levels (Wilson, 1997). Nevertheless, it has been shown that this new approach to group selection is just a different way of splitting up the process of natural selection than kin selection theory (Hamilton, 1975; Grafen, 1984; Wenseleers *et al.,* 2010), and that it's predictions are always exactly identical. This is because in the case of group selection, altruistic traits can only be favoured if there is sufficiently high between-group genetic variance in this

trait, which from a kin selection angle is only the case if within groups individuals are genetically related.

### *3.2 Direct reciprocity*

When looking at human society, other species, and even across species, it becomes clear that cooperation does not only take place between related individuals. To explain this, Trivers (1971) put forward the idea of reciprocal altruism, suggesting that individuals who repeatedly interact can take turns in helping each other. Of course, such reciprocal behaviours are only altruistic in the short term: fitness benefits should be obtained during a person's life span for any reciprocal helping to be able to evolve. By consequence, 'direct reciprocity' is preferred as a label for these repeated interactions (Alexander, 1974; West *et al.,* 2007b). Every interaction leaves the interactants with a choice of strategy: will they cooperate or defect? Such situations are also known as repeated Prisoner's Dilemmas in game theory (we will elaborate more on this in I. 4.2.2). In this game, the selfish option of defecting provides a higher payoff than cooperation regardless of what the other person does. Yet a situation in which 2 people cooperate earns them both a higher payoff than a situation in which 2 people defect, which means that cooperating might be fruitful (Axelrod and Hamilton, 1981). As with kin selection, the evolution of cooperation is only possible when a certain factor exceeds the cost-to-benefit ratio of the altruistic act. In the case of direct reciprocity, that factor has been shown to be w (Henrich and Henrich, 2007), which is defined as the probability of another encounter between the same two individuals.

The remark can be made that direct reciprocity is not free of constraints: there must be some mechanisms for recognition and discrimination of previous interactants, for determining the next action as a function of the previous interactions and for estimating the probabilities of future interactions. There is evidence for reciprocal interactions in non-human primates, for instance grooming (Silk, 2005). This fuels the already circulating question to what extent these constraining conditions have to be met in order for direct reciprocity to operate. Additionally, other studies point out that humans as well as other animals have a more implicit, emotional capacity for probability estimation (Loewenstein *et al.,* 2001).

### *3.3 Indirect reciprocity*

Kin selection and direct reciprocity both still fail to explain that people sometimes offer help to strangers in need or donate money to a charitable cause. Offering help does not always result in a reciprocated action, let alone you could foretell for sure if you will meet someone again in the future. All things being equal, the formation of a reputation is a mechanism that allows for future reciprocal services even though your direct interaction partner is not reciprocating. Being an exceptional cooperator can earn you the good reputation that might encourage others to provide you a 'reward' later (Alexander, 1985; Boyd and Richerson, 1989). Indirect reciprocity requires high levels of information processing (for image scoring) and a capacity to pass on this information (language) and is not often described in non-human species. Therefore it is sometimes thought that this mechanism has played a decisive role in the evolution of intelligence, morality and social norms, setting us apart from other species. Again a simple rule can be established: indirect reciprocity can only enable the evolution of cooperation if the probability of knowing someone's reputation is bigger than the cost to benefit ratio of the trait (Henrich and Henrich, 2007).

Several have noted that indirect reciprocity can only be an evolutionary stable mechanism in small groups (Boyd and Richerson, 1989). However, an equilibrium is established when the display of costly actions is followed by a pair-wise cooperative interaction (Roberts, 1998). This can be categorized under another conceptual evolutionary framework, referred to as costly signalling theory (Smith and Bliege Bird, 2005).

### *3.4 Costly signalling*

Costly signalling theory arose from the observation that some creatures incur such great costs on displaying some behavioural or morphological trait while, from an evolutionary perspective, it seems gobsmacking that such a 'handicap' could have evolved (Zahavi, 1975). One theory is that these expensive traits are signals conveying honest information about the signaler's qualities as a potential social interactant, where honest information requires that the cost experienced by the signaler is linked to the quality that is being advertised. A quality is a trait that is usually difficult to assess directly, like good health, solidarity, leadership ability or commitment to an on-going cause (Smith and Bird, 2000). For costly signalling to be maintained by selection, both signalers and interactants must benefit from this sharing of information. Applied to cooperation, for example, an act of cooperation

could provide group members with a reliable signal about a quality of the signaler and by consequence those group members might prefer them, for purely selfish reasons, as a mate or an ally. An important difference with indirect reciprocity is thus that the benefits delivered by costly signalling do not necessarily have to be the result of a subsequent pair-wise interaction: the benefits might be obtained from avoiding conflict with the signaller or through the display of the signal itself. This framework also offers insight on why we take such interest in gossip: we want to discover who it might be beneficial to interact with and also who we have to avoid to all extent (Smith and Bird, 2000).

### *3.5 Punishment*

Both theoretical and experimental work point out that punishment of selfish individuals is a potent mechanism for promoting the evolution of cooperation in humans (Boyd and Richerson, 1992; Fehr and Gachter, 2002; Sigmund, 2007). Punishment is also a focus of interest in the study of other social animals and it has been shown that this type of negative reciprocal behavior occurs frequently and for various numbers of reasons. One of those reasons, which is of special interest to us, is to coerce cooperative behavior (Clutton-Brock and Parker, 1995). Recent findings even indicate that inclusive fitness theory on its own is not a satisfactory explanation for the (sometimes extreme) levels of altruism observed in many modern insect societies and that enforcement is a key factor in maintaining this altruistic behavior (Ratnieks and Wenseleers, 2008). So generally, punishment is considered 'prosocial' because it facilitates evolution of cooperation. Public goods games also reveal the existence of antisocial punishment, which is the punishment of cooperators (Rand and Nowak, 2011). In this literature study, we chose to further elaborate on theories of prosocial punishment.

That being said, punishment requires time, energy and risk, and thus the ultimate question "how does such costly behaviour evolve?" still lingers. In dyadic interactions, the punisher is sole beneficiary but in larger groups, those benefits are shared by others who didn't pay the costs. Hence, punishment poses a second-order public goods problem that increases as the costs of punishing others get higher (Boyd *et al.,* 2003). In section I. 5, we discuss prosocial punishment to a much larger extend and we try to give an overview of the proposed explanations to the second-order public goods problem.

# 4 Experimental and behavioural economics

While game theory provides us with the mathematical language for describing strategic interactions between players and making predictions about the outcomes of those interactions as well as the emergence of Nash equilibria, the predictions that followed those mathematical models were still to be tested in the field. As a consequence, economists began to run laboratory experiments, using undergraduate students as subjects. From the early 1980s on, a number of experimental games were developed, of which we will give a few important examples in section 4.2. A lot of the research done in behavioural economics involves 'revealed preferences', a term used to refer to the choices people make.

## *4.1 Experimental game theory*

Typically, experimental game theory experiments entail economic decision-making with real, often substantial, monetary stakes (Camerer and Fehr, 2004). There are some other standard experimental conditions: subjects are anonymous, only play once and cannot communicate with other people. The description of the game usually stays quite abstract, with numbers and letters being used to represent strategies rather than elaborate strategy descriptions (Camerer and Fehr, 2004). Such conditions are not claimed to represent lifelike situations and therefore require careful interpretation (Levitt and List, 2007), but they provide a baseline for investigating the effect that certain factors have on the players' strategic decisions. For instance, framing effects can be looked into, since players usually behave differently as a distinctive context is created by the game's descriptions. Also, economists are very interested in finding stable strategies (the Nash equilibrium), so games will commonly be played repeatedly to allow for learning and equilibration to occur. According to standard procedure, subjects are usually asked to answer some comprehensive questions concerning their payoff calculations before the start of the game. Experimental economists insist on actually paying their subjects' earnings from the game, plus a small show-up fee.

## *4.2 Examples of games*

In the following subsections, we present three common examples of economic games: the ultimatum game, the prisoner's dilemma game and the public goods game.

### 4.2.1 Ultimatum game

The ultimatum game is played with two players who have to agree on the division of a sum of money. One player gets assigned the role of the proposer and has to decide what part of that sum he wants to keep for himself and what part he wants to donate to the other player. This second player is the responder and has the opportunity to either accept the offer or reject it. When the offer is accepted, both players get paid accordingly. When it is rejected, both players get nothing. The ultimatum game is played anonymously, so neither player has a reputation to gain or to maintain. In the standard view of game theory, the rational solution of the ultimatum game is that the proposer will offer the responder a very small fraction of the money, while the responder will take whatever he can get since getting some money would be preferred over getting none at all (Güth *et al.,* 1982). However, empirical data obtained from such games undermines this prediction: proposers mostly offer up to 50% of the available sum and responders often reject low offers, with half of the responders rejecting shares of less than one-third of the sum (Thaler*,* 1988; Nowak *et al.,* 2000). Biologically speaking, these games could represent situations in which individuals try to agree on the future division of a reward of cooperative hunting, or the formation of an alliance, or a dilemma of food sharing (Page *et al.,* 2000).

### 4.2.2 Prisoner's dilemma

The prisoner's dilemma game reflects a situation in which the 2 persons that are interacting both have a choice to cooperate or to defect (Rapoport and Chammah*,* 1965). The payoff matrix of this game is given in table 1. In the prisoner's dilemma, it is always individually the best strategy to defect: given that your opponent cooperates, you get the highest possible payoff (T) by defecting while your opponent gets the lowest possible payoff (S); if your opponent defects, you earn more by also defecting (P) than by cooperating (S). However, if both players defect their payoffs (P) will be lower than when both players would have cooperated (R). This becomes clearer when considering the payoff matrix of this game, of which the payoffs fulfil the following relations: $T > R > P > S$ and $R > (T+S)/2$. In this game, mutual defection is the only Nash equilibrium.

**Table 1: Payoff matrix of the prisoner's dilemma game. 'P' stands for punishment, 'R' stands for reward, 'S' stands for sucker and 'T' for temptation.**

| Person 1/person 2 | Cooperate | Defect |
|---|---|---|
| Cooperate | R, R | S, T |
| Defect | T, S | P,P |

### 4.2.3 Public goods game

Much of human cooperation involves large numbers of individuals and public goods games supply an appropriate tool for studying such *n*-person interactions, including prosocial behaviour in a group context (Camerer and Fehr, 2004). The public goods game structure is similar to that of the prisoner's dilemma in the sense that individually, a given player is best off not to contribute anything to the public good, whereas from the group's perspective, contributing all ones money would be the most beneficial (Hardin, 1971). This situation results in the problem of common goods: when mechanisms to provide a direct or indirect fitness benefit (section 1.3) to the cooperator are absent, cooperation is not a stable outcome of the game. The game itself follows a fixed set of steps: each of the *n* group member get a number of tokens (these could be units of money) and they can freely choose whether and how much of this endowment they want to invest in the group's public good. The money invested in the public good is multiplied by a factor $r > 1$ and then evenly redistributed over n group members. For each token invested in the common good, each person gets a share $\alpha$ ($0 < \alpha < 1$). Since every player benefits from the investments of others, free riding is possible. After all, if a subject keeps all of his endowment $x$ whilst the other subject contribute their whole endowment to the public good, then this defector will get the total sum of $x + \alpha$ while the cooperators only get $\alpha$.

Both the prisoner's dilemma and the public goods game are representative of situations in which, for example, a depletable resource has to be shared among a group of people, while free-riders cannot easily be excluded from sharing. The shared good could be a number of things, such as clean air, fresh water or common fishing grounds (Ostrom, 2000). This dilemma over the use of common resources is often referred to as 'the tragedy of the commons' (Hardin, 1968) and describes how

people will ultimately deplete a common resource because of selfish motives: they refuse to limit their consumption of the resource because this provides them with immediate benefits, while in the long run this behaviour actually yields the worst outcome.

## 5 Altruistic punishment

Cooperation is a costly action that benefits others and we already mentioned how, historically, this has been a hard nut to crack for many biologists. Researchers have summoned upon many theories to make evolutionary sense out of this behaviour (I. 3), but modern society still continues to baffle us. Cooperation levels are very high and the tragedy of the commons isn't always realized. When defection yields the largest personal payoff, involving punishment can drastically change the outcomes of interactions: If punishment of non-cooperators is possible, then this in itself can be an incentive for individuals to cooperate so as to avoid the costly consequences of a penalty (I. 3.5). In our society, social norms are enforced by laws, free- riders are discouraged by government institutions that implement taxations and by informal mechanisms like ostracism of cheaters. Such established mechanisms can provide a remedy for a lack of voluntary cooperation (Camerer and Fehr*,* 2004; Sigmund*,* 2007).

In order to investigate whether punishment would raise cooperation levels in public goods games, Fehr and Gächter (2000; 2002) conducted a public goods experiment with the opportunity for punishment. Subjects remained anonymous throughout the experiment and were never matched with the same players in the different trials. Participants were well informed about these anonymous, one-shot conditions, so Fehr and Gächter argue that the game is absent of the selfish motives of reputation building and reciprocity. It turns out that the grand majority of punishment in the laboratory experiment was executed by cooperators and imposed onto defectors. The latter are defined respectively as above-average contributors and as below-average contributors. Their results also indicate that punishment significantly increases cooperation levels and can maintain cooperation under conditions in which pure selfishness would otherwise lead to an inevitable breakdown of cooperation (Fig. 2). Fehr and Gächter (2002) label the subjects who punish as 'altruistic punishers': they impose penalties on free riders even though this is costly and yields no material gain for themselves, whereas the future group members of the punishee could potentially benefit from the punishment that was executed. This benefit could be obtained if the punished subject reacts to the penalty by raising its

contributions in the next rounds as to avoid future acts of punishment. As mentioned in section I. 3.5, however, punishment also leads to second-order public goods problem in the sense that punishment itself is costly to the actors. Fehr an Gächter (2002) claim that this second-order problem is solved “if enough people have a tendency for altruistic punishment”. It is clear, however, that this statement does not offer an ultimate explanation for why altruistic punishment would occur and how it could evolve. In the next section, we will situate altruistic punishment as a part of a broader behavioural tendency that has been put forward by game theoreticians, namely ‘strong reciprocity’. In section I. 5.2, we will then start providing possible explanations for the occurrence of altruistic punishment.

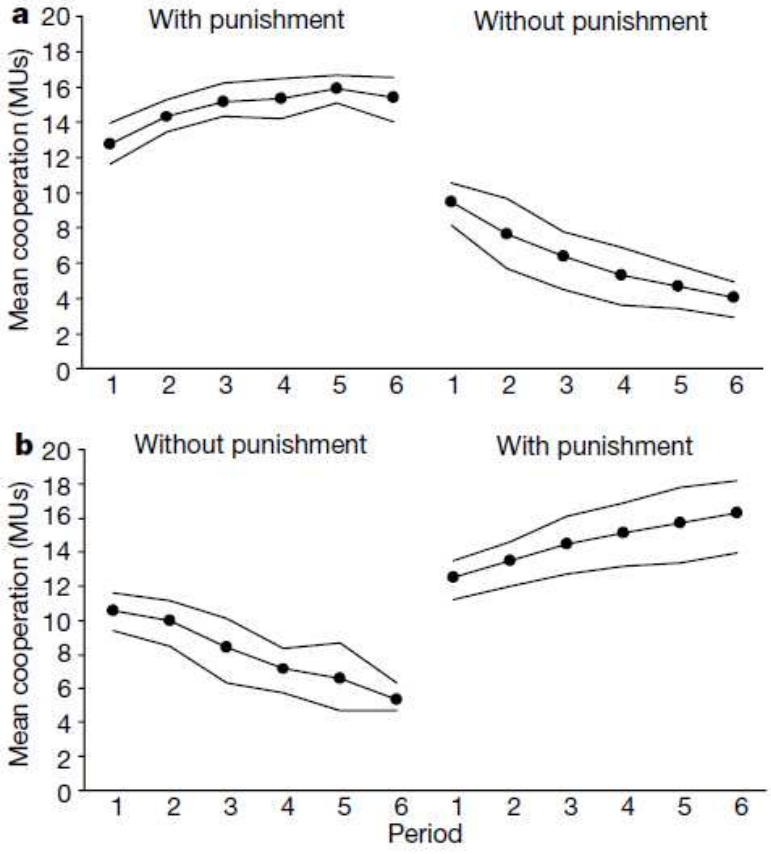



**Figure 2: The effect of punishment on cooperation levels in Fehr and Gächter’s experiment (2002).**

### *5.1 Strong reciprocity and morality*

Strong reciprocity is used as a more inclusive expression for a behavioural strategy that involves the willingness to reinforce cooperative behaviour by reciprocating accordingly, even if this is costly and doesn't offer any (present or future) rewards to the reciprocator (Fehr *et al.,* 2002). Fehr and Henrich (2003) define this in a clear manner: "A person is a strong reciprocator if he is willing (i) to sacrifice resources to bestow benefits on those who have bestowed benefits (= strong positive reciprocity) and (ii) to sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm (= strong negative reciprocity)".

Since the focus of this thesis is on strong negative reciprocity, it must be mentioned that people don't only punish those who have treated them unfairly in a one-shot game, but also those who have treated others unfairly (Fehr *et al.,* 2002; Gintis *et al.,* 2008). These experimental findings oppose the view, shared by the behavioural sciences and much evolutionary thinking, that all behaviour should ultimately be self-interested. Gintis (2008) argues that this strengthens the belief that strong reciprocity has been the missing puzzle piece in explaining high levels of cooperation and altruism in humans, as it supplies a powerful device for enforcing norms that prescribe, for example, food sharing or collective action (Fehr *et al.,* 2002).

Strong reciprocity is put forward as one of the building blocks of human morality, along with other-regarding emotions such as empathy, shame and envy and personality traits such as honesty, trustworthiness, and others (Gintis *et al.,* 2008). However, there is an on-going discussion concerning the adaptiveness of strong reciprocity and human morality which we will further elaborate on in section I. 5.4.1 (Gintis *et al.,* 2008; Price*,* 2008).

### *5.2 Proximate mechanisms*

Fehr and Gächter (2002) use hypothetical investment scenarios to elicit the underlying reasons for punishing. They hypothesize that the particular pattern of punishment that is observed, namely co-operators punishing defectors, may be explained at the proximate level by an analogous emotional pattern. The results point out that altruistic punishers indeed experience negative emotions towards the subjects that free ride on their contributions in the public goods game. Moreover, most people expect these emotions, which explains why the threat of punishment alone has an immediate positive influence on cooperation levels (Fehr and Gachter*,*

2002). Neuroscientific evidence for emotional patterns associated with punishment comes from the recently emerging field of neuro-economics.

In a neuroscientific study, de Quervain et al. investigated the neural basis of altruistic punishment by scanning the subjects' brains while these subjects learned about the defectors' unfair behaviour and decided on what punitive measures to inflict on the defectors (de Quervain *et al.,* 2004). They were able to pinpoint the activation of the anterior dorsal stratium, in which the caudate is located. This is a brain region associated with the making of decisions or taking of actions that are motivated by anticipated rewards. Moreover, they discovered a positive correlation between peoples willingness to pay a higher cost of punishment and activation of the dorsal stratium. This neuroscientific result reinforces the view that people punishing those violating fairness norms are proximally driven by emotions. In this case, the satisfaction that they get out of 'justice being served' seems to motivate acts of punishment.

### *5.3 Ultimate explanations*

Given that people seemingly evolved proximate mechanisms for the punishment of free-riders, why would the evolution of strongly reciprocal behaviour be favoured if there is no personal gain to be obtained? In the subsequent sections, we attempt to give an overview of the ultimate explanations that have been given for altruistic punishment. Although this thesis mostly deals with punishment, sometimes more inclusive terms such as 'strong reciprocity', 'other-regarding utility function' or 'internal altruistic norms' will be used instead of 'altruistic punishment'. Please note that this is not to imply that these expressions are synonymous to each other. Rather, this is a consequence of the fact that the search for ultimate explanations of these concepts often runs in parallel in the literature. For the sake of simplicity and correctness, we use the same designations as in the literature that we cite.

#### 5.3.1 Inequity averse utility functions

Classical game theoreticians assume that preferences are self-regarding and outcome oriented. This assumption is reflected in the utility function, where the outcome yielding the biggest payoff corresponds with the biggest utility. However, empirical evidence emphasises that people do attach importance to a fair division of the goods and the previously introduced form of the utility function does not take into account any social preferences such as fairness or equality (Fehr and Schmidt,

1999). This lead to the development of an inequity averse utility function, which gives rise to a new concept of rationality: People act rationally, not simply by maximizing their own outcome, but by pursuing a minimal difference between payoffs of both players engaged in the game (Fehr and Schmidt, 1999). However, this new utility function again shifts the problem: Human preferences are thought to be 'other-regarding' and utility is redefined so that theoreticians did not have to drop their assumption of human rationality. Yet even if utility now takes into account social or cultural preferences, we still have to consider what adaptive benefits these preferences confer.

### 5.3.2 Cultural group selection

The basic idea of group selection was introduced in I. 3.1 and states that altruistic traits might evolve through differential extinction rates of groups with different numbers of altruists due to inter-group conflicts. For example, cooperative groups might be better at warfare or dealing with food resources. Even though group selection has historically proven to not be devoid of pitfalls, theoreticians still turn to this idea in order to ultimately explain strong reciprocal behaviour. Several authors (Bowles and Gintis, 2003; Boyd *et al.,* 2003; Fehr and Fischbacher, 2003) have involved the concept of cultural group selection to explain the evolution of altruistic punishment and 'other-regarding' utility functions (I. 5.3.1). Cultural traits can be transmitted in ways that are distinct from those of genes and such cultural processes play a crucial role in the evolution of human behaviour.

We want to remind the reader that punishment creates a second-order problem: Having altruistic punishers in the group might make cooperation a less costly strategy than defection, but this cooperative state will never hold, since punishers will be outcompeted by co-operators who don't bear the extra costs of punishing. Boyd, Gintis et al. (2003) believe that the puzzle of altruistic cooperation can truly be solved by the involvement of altruistic punishment because there is a key difference between these two phenomena: The relative payoff disadvantage of a co-operator is not dependent on the number of defectors in the group, while the relative cost of being an altruistic punisher is. When altruistic punishers are common in the group, defectors become scarce because of the effective punishment, which in its turn ascertains that acts of punishment are rarely needed anymore. So in effect, 'many hands make light work' for altruistic punishers. Group level selection will create great benefits for groups with a lot of altruistic punishers and, on the level of the individual punisher, only a very small disadvantage. Henrich and Boyd developed an

evolutionary model where strong reciprocal norms are adopted through payoff-biased transmission (imitation of the most successful individuals in the group) and conformist transmission (imitation of high frequency behaviour in the group) (Henrich and Boyd, 2001). Boyd, Gintis et al.'s (2003) cultural group selection model includes payoff-biased transmission that maintains the between-group differences needed for group selection to work and shows how punishment can be maintained when common. A remaining challenge, however, is to explain how a strong reciprocator strategy can invade into a population initially consisting only of defectors.

#### 5.3.3 Hitchhiking of altruistic norms

If you think of a person with good moral values, you will probably imagine someone whose altruistic behaviour is not dependent on some external threat of punishment or reward. The adjective 'good' will rather reflect the fact that this person *wants to* behave altruistically (Gintis, 2003). This behaviour is due to 'internal norms', which are cultural norms that are enforced from within the individual by feelings such as guilt, shame or responsibility. At some point in man's evolutionary past, we must have gained this ability for internalization of norms. Gintis (2003) proposes a model in which the capacity for internalization of norms is given by an allele at a specific genetic locus, while different norms only manifest themselves on a phenotypic level. He claims that this genetic trait is overall fitness-enhancing because it gives humans the advantage of rapid cultural adaptation. If there is a selfish internal norm that ensures a sufficiently big increase in fitness, it can invade a population of normless phenotypes. Costly altruistic norms can then 'hitchhike' on the trait's general beneficial characteristics: they can become internalized as well, without the genetic trait of norm internalization being selected against. From here on, cultural group selection is thought to ensure the maintenance of prosocial internal norms through differential group success. To conclude his argument, Gintis also points out that norms prescribing behaviour that is much alike personally fitness-enhancing behaviour, would be easier to internalize. It seems plausible that at first, punishing defectors was beneficial through reputation building, and then later on became an internalized moral principle.

### *5.4 Opposition to prevailing explanations of altruistic punishment*

As far as ultimate explanations of altruistic punishment are concerned, it is clear that there is still no consensus in the literature as to what represents the most likely

theory. In this section, we present three alternative schools of thought that question the previously introduced theories.

### 5.4.1 Misfiring hypothesis

Earlier on, we briefly introduced evolutionary psychologists' view on the human brain and our behaviour. They consider the selection process as the optimizing engine for behaviour, rather than the brain itself, and argue that our bodies house a stone age-mind (I. 1.1). Since our minds are executors of adaptations, rather than maximizers, they will not be adapted to certain aspects of the novel environments modern society and laboratory experiments expose them to. As a consequence, our brain sometimes 'misfires' and some of our human behaviour will be maladaptive in the specific experimental conditions that researchers place their subjects in. Although researchers that believe in the misfiring hypothesis do not deny the presence or importance of strong reciprocity and moral norms in everyday life, they offer a critical standpoint on the context that is created in economical experiments. They interpret strong reciprocal behaviour, and human morality in general, with respect to the relevant interactions ancestral evolutionary environment and find that the big discrepancy between this environment and that of the experiments lies in the prevalence of anonymous, one-shot encounters (Gintis *et al.,* 2008; Price*,* 2008).

#### *5.4.1.1 The artificiality of one-shot, anonymous encounters*

The main focus in the misfiring theoreticians' argumentation concerns the anonymous, one-shot conditions in experimental games. They believe that these conditions constitute such an artificial environment that this causes the brain to misfire, as in man's evolutionary past, both interactions with strangers and anonymous interactions would have been very rare.

**a) One-shot encounters**

A prime argument of many evolutionary psychologists is that in hunter-gatherer societies, encountering a stranger was a very rare event. Therefore, our Pleistocene brain is not well adapted to interactions with people other than close acquaintances or kin. Researchers who treat strong reciprocity as an adaptation argue that there is ethnographic evidence for one shot encounters (Fehr and Henrich*,* 2003). However, opponents respond to their evidence in three ways (Hagen and Hammerstein*,* 2006). First, in case of such an encounter, how can it be known if the probability of future encounters will be zero (or near zero)? For example, when one group raids another, the prospect of a counter-raid would not seem unrealistic. The second concern

unveils a more basic problem with economic games in general. Basically, even if one-shot encounters were common enough in our evolutionary history to allow an adaptive response, it would be less than obvious that our brain would never misfire in a game situation, which is always artificial. The third response is probably the most destructive and shows a contradiction between the ultimate explanation of strong reciprocity and its proximate description. The proposed model for the evolution of strong reciprocity and these other-regarding norms is a (cultural) group selection model, implying that this behaviour is selected for as within-group behaviour by between-group selection. However, if we assume one-shot encounters were common in the evolutionary past, almost by definition this would have took place between groups. This seems to leave the followers of the strong reciprocity theory with a gap in their argumentation. Additional knowledge, such as whether the players in the game regard their fellow players as in-group or out-group, becomes crucial (Hagen and Hammerstein, 2006).

**b) Anonymous encounters**

Likewise, the concept of real anonymity raises a big question mark. Researchers that believe in the misfiring hypothesis presume that anonymous interactions and encounters were extremely rare and that there would always be some incentive for reputation formation. Our psychological mechanisms must therefore be adapted for situations in which reputation is at stake, hence our limited brains would never really consider a situation in which we are not being watched and evaluated (Price, 2008). Hard ethnographic evidence for anonymous interplay is indeed lacking, even though it has been shown that people behave differently in anonymous relative to non-anonymous games (Fehr and Henrich, 2003). This pro-adaptation argument is also applicable to the discussion on one-shot interactions: people can (emotionally) distinguish between strangers and partners and seem to grab the meaning of these one-shot circumstances. Yet the new question can be posed how penetrable these underlying emotions are (Hagen and Hammerstein, 2006). For example, a man can get an erection from a picture in Playboy magazine. Even though he surely grabs the fact that this woman is not there with him in the flesh (so his erection is quite pointless), he nevertheless gets excited. Hence, in this case, it is clear that the underlying, cognitively impenetrable mechanism is not triggered correctly, and the same might well apply for experiments using anonymous one-shot interactions.

#### *5.4.1.2 Misfiring in other animals*

Strong reciprocity theorists hold another argument against the misfiring hypothesis: The brain of non-human primates does not appear to misfire. In other words, their

brain does not trick them into displaying strong reciprocal behaviour, even though they can also experience 'artificial' environments analogous to human modern societies (Fehr and Henrich, 2003). Indeed, in a study where a mini-ultimatum game is played with chimpanzees (*Pan troglodytes*), our closest living relatives, it was shown that chimpanzees display rational behaviour (Jensen et al., 2007). Chimpanzees were found to act according to traditional economic models of self-interest, generally accepting any above-zero offer, regardless of the offer being fair or not. These results are contrary to those of similar experimental games played with humans, in which it was found that people behave according to fairness and other-regarding norms. However, we believe that this is no more an argument for strong reciprocal behaviour being an adaptation than it is for being a maladaption.

To conclude the misfiring debate, we would like to point out that the interpreting of experimental games should always be done with caution. Nevertheless, even games with anonymous, one-shot interactions can offer insight into the proximate mechanisms that control human behaviour.

### 5.4.2 Criticism of cultural group selection models

Without claiming to be complete, we will introduce two important sources of critique on cultural group selection as an ultimate account for altruistic punishment. The first one concerns the phenomenon of 'parochial altruism' and draws attention to the fact that altruistic norms are not confined to ones' own group. The second line of critique fundamentally questions whether culture really alters or speeds up the evolution of strong reciprocity.

If one follows the group selection reasoning, norm violations should be negatively enforced only within groups, since the group can only benefit from the individually costly punishment behaviour if it is used to establish an internal, group-beneficial social norm. This norm could for example apply to food sharing, collective hunting, participation in warfare, or other traits that can make the group come out on top in inter-group conflicts (Bernhard *et al.,* 2006). Such predictions, however, are not supported by Bernhard, Fishbacher et al.'s experiments, which instead demonstrate that human altruism follows a parochial pattern. Parochialism is defined as a preference for in-group members and that group could be defined for example on one's ethnicity, race or language. This is nicely illustrated in the experiment, where a dictator gets to divide a sum of money between himself and a receiver, and a third person gets to decide whether the dictator should be punished or not. Parochialism

manifests itself as the third person that punishes in- as well as out-group members if the person that was treated unfairly is an ingroup to this third person. Thus, certain social norms seem to extend the boundaries of one's own group. These findings point out the lack of consideration of other factors in cultural group selection models. By punishing outsiders who harm an ingroup individual, the group might create a 'don't mess with us'-reputation that increases their overall security by preventing attacks.

The second source of critique on cultural group selection models comes from Lehmann et al. (2007), who argue that it is not clear how cultural transmission could lead to different selective pressures on the evolution of strong reciprocity than traditional genetic evolution. In support of this, Lehmann et al. constructed a mathematical model to thoroughly investigate which conditions would allow the invasion and evolution of a helping strategy and punishing strategy (strong reciprocity), focusing on the effect of spatial structure (limited dispersal), linkage of both traits (strong reciprocity as a single Mendelian trait) and different modes of cultural transmission. Importantly, they found that different types of cultural transmission cause the selective pressure on strong reciprocity to increase, decrease or not change at all compared to that under genetic transmission. They also conclude that punishing non-cooperators cannot be favoured unless the two traits are linked. Finally, they remind us that even if cultural group selection is active, this kind of evolution would not favour strongly reciprocal behaviour towards strangers (I. 5.4.1.1). The model results indicate that strong reciprocal behaviour will evolve through mechanisms of genetic or cultural kinship, and that the concept of strong reciprocity should not be misinterpreted as an ultimate mechanism in itself but that instead it should still ultimately have a purely selfish basis. Their results tend to support the view that punishment would quality as a spiteful behaviour, since punishment in their model is carried out to reduce competition between group members (induce fitness costs). In addition, their results indicate that the critiques (Hagen and Hammerstein, 2006; West *et al.,* 2010) that have been made on the interpretation of experimental games can in no case be neglected. Boyd, Richerson et al. (2011) respond to Lehmann et al. (2007), stating that the authors reach such conclusions by making different and less realistic assumptions in their model. It is thus essential to find out what assumptions best fit the empirical data about human learning, cultural diffusion and human cooperation. Lehmann et al. assume that adaptive forces in cultural evolution are weak compared to migration. However, empirical evidence of cultural transmission in humans have shown that considerable

cultural changes can occur and go to fixation in sometimes less than one generation (Boyd *et al.,* 2011).

### 5.4.3 Selfish punishment

Up until this point, the kind of prosocial punishment that we discussed was considered altruistic, supposedly arising from man's altruistic, inequity averse nature that motivates his behaviour. It is not surprising that one would draw such conclusions, since experiments show subjects who engage in costly punishment only to benefit others. Eldakar et al. (2007), however, unveil another possible incentive for prosocial punishment. Their results, gained from model simulations and fictional scenarios, indicate that cheaters also engage in costly punishment, affecting other cheaters. Eldakar et al. name this phenomenon 'selfish punishment'. In contrast with altruistic punishers, a selfish punisher acts selfishly in the context of the first-order public good (that of cooperation) and acts altruistically only in the context of the second-order public good (that of punishment) (Eldakar and Wilson*,* 2008). The motive for a cheater to punish 'one of his own' is quite straightforward: a cheater's fitness will increase when other group members contribute more to the public good. If everyone would start cheating, not enough altruists would be left to exploit. Cheaters can undermine each other in yet another way: the probability that cheating is detected by the group members gets bigger when there are more cheaters present in the group, which increases the probability that the cheater will get punished. In brief, defectors are in constant competition with each other and are therefore inclined to punish (Eldakar *et al.,* 2007).

### 5.4.4 Theoretical models on selfish punishment

In Eldakar et al.'s (2007) model, altruism is considered directly proportional to the part of the endowment allocated to the group fund. Throughout the simulation runs, both the parameter for altruism and the propensity to punish (P) are varied from 0 to 1 in all possible combinations to see what correlations arise between altruism and punishment. Their model shows a stable equilibrium between altruistic non-punishers and selfish punishers. They also varied a number of other parameters and concluded that the cost of punishment presumably plays an important role in the evolution of punishment and cooperation: when the cost is small, the correlation between punishment and altruism is close to zero, however this correlation becomes increasingly negative as the punishment cost increases. These findings reflect the fact that, unlike altruistic punishers, selfish punishers are able to compensate for the

cost of punishment through their selfish behaviour in the first-order public good (by taking advantage of the cooperators in the group) (Eldakar *et al*., 2007). Such a situation can be regarded as a division of labour (Eldakar and Wilson*,* 2008). Thus one would expect selfish punishment to be a very relevant concept, especially if punishing cheaters is costly. Nakamaru et al.'s (2006) models show that under certain conditions, selfish punishment promotes the spread of an altruistic punisher strategy. By consequence, selfish punishment could provide an ultimate explanation for the evolution and maintenance of high levels of altruism, whereas according to their models, altruistic punishment in its own cannot (Nakamaru and Iwasa*,* 2006).

## 6 Aims of the thesis and hypotheses

In the existing literature, punishment is put forward as one of the driving forces behind human (and non-human) cooperation because it can be used to impose costs on those who don't cooperate (section I. 3.5, I. 5). However, the focus of the research on prosocial punishment has mainly concerned altruistic punishment, a strategy that has proved hard to ultimately explain. Since other strategies and interactions between strategies have been mentioned in the literature (Eldakar et al., 2007)((Rand and Nowak*,* 2011)((Nakamaru and Iwasa*,* 2006), we want to further investigate punishment. To do this, we will conduct an anonymous, one-shot, optional public goods game with a manipulated opportunity for punishment and cost of punishment. In analogy with Fehr and Gächter (2000, 2002), we hypothesize that punishment will significantly raise the average contribution level. From the same literature, we derive the hypothesis that the chance that a person will punish his interactant will be positively correlated to the actor's degree of altruism (contribution). In particular, we will focus on the possibility that other punisher strategies are also present. We hypothesize that a significant part of prosocial punishment will be executed by selfish punishers as a strategy to reduce competition between(a) defectors or (b) defectors and loners (Eldakar, Farell et al. 2007; (Rand and Nowak*,* 2011). This would lead to a selfish, hypocritical vision of the person who punishes rather than a purely altruistic, morally responsible punisher, like Fehr suggested. We will also investigate Eldakar's related model predictions (Eldakar, Farell et al., 2007) concerning punishment costs. In accordance with those predictions, we hypothesize that as the costs of punishment increases, the amount of prosocial punishment that can be accounted for by a selfish punishment strategy will increase. We also postulate that punishing will not be the winning strategy (Dreber *et al.,* 2008).

Participation in our public goods game is optional, which means that any participant can choose to either play the public goods game, or opt out for a fixed sum.

We are interested in investigating scenarios where cooperation is optional for a number of reasons. First of all, not that much research has been done using optional public goods games. However, when considering human's evolutionary past, it seems plausible that in many scenarios there are more options than just defecting and cooperating. This could be the case with hunting on big-game (Gintis*,* 2005), where one could choose a safe source of income by hunting alone, on smaller prey, instead of joining a hunting party. These non-participators or 'loners' are considered risk-averse: they neither take the chance of ending up empty-handed in case the risky hunt is unsuccessful, nor do they risk being cheated on by defectors in their hunting group, even though the potential personal gains of hunting big-game are substantial. It is important to remark that the difference between a loner and a defector is that the loner neither receives any benefits nor pays the costs of the common good (Hauert *et al.,* 2002; Fowler*,* 2005). A second, but equally important, reason is that sometimes an optional game turns out necessary to make sense of the observed behavior. In their study, Rand & Nowak (2011) concluded that the antisocial punishment they observed in compulsory public goods games was carried out by defectors. However, in the optional version of the game, these antisocial punishers predominantly chose a loner strategy, which led to the conclusion that loners are the ones inclined to punish cooperators, not defectors. The observed antisocial punishment strategy of loners is an ESS according to the model presented in their paper (Rand and Nowak*,* 2011). This example raises the suspicion that carrying out a compulsory public goods game could lead to a distorted image: a considerable part of defectors (that would choose a loner strategy if available) could be involved in the punishment of cooperators, instead of in the punishment of defectors or loners. In our experiment, such 'hidden' effects might alter the degree to which selfishness (defecting) in the first round and punishing in the second round are correlated, since loners, if given the option, would not have participated in the game and would have accounted for neither being selfish or altruistic punishers. The third argument for conducting an optional public goods game involves the information on the role of loners in prosocial punishment. Following Eldakar, Farell et al.'s (2007) prediction on antisocial punishment, we hypothesize that loners would not be involved in the punishment of defectors, yet would be involved in the punishment of cooperators.

# II. Materials and methods

First, we provide the reader with the specifics of the experimental setup. We then summarize the course of the game by offering an overview of the steps that a participant in the game had to go through. Next, we supply some information on how the experiment was programmed and finally, we describe the methods used for data analysis.

## 1 Experimental setup

A total of 96 persons from the database of the marketing research group of the KULeuven were gathered for this experiment, all of which were students. The experiment consisted of 6 sessions with 16 participants per session. The optional public goods game itself was mediated by a computer program and all game interactions thus took place through this program; instructions and other players' decisions were shown on the participants' computer screens. The sessions took place in 4 separate computer labs and every lab room housed 4 participants that were each in a different corner of the room, to make sure their game decisions were invisible to the other participants in the room. Participants received either a show-up fee of 5 euro, or their average payoff earned in the game in case this payoff exceeded this 5 euro.

Every session, 16 persons played an optional public goods game made up of 12 rounds. Those rounds were evenly distributed over the 3 different experimental conditions (4 rounds/condition): a control condition, where a public goods game was played without opportunity for punishment; a first treatment condition with punishment opportunity where the cost of punishing was low (1:3 ratio, which indicates that the punishee pays three times as much as the person that punishes); and a second treatment condition where the cost of punishing was high ( 2:3 ratio). Practically, this implies that in the low cost condition, punishing one person cost 2 euro while in the high cost condition, this cost was 4 euro. In both conditions, the punishee lost 6 euro per punishment that he or she received. From here on, we will refer to the first treatment condition as the low cost condition and to the second treatment condition as the high cost condition.

Every session, all 16 players got to play 4 rounds in all 3 experimental conditions.

Every round, 4 new groups of 4 participants were formed. Before the start of the experiment all participants were informed that the game was played anonymously.

Also, in one and the same condition all interactions were one-shot, meaning that no two players would interact with each other more than once. Between conditions, the participants' player numbers were randomly redistributed so that no reputation formation was possible.

### *1.1 Game course*

At the start of the experiment, subjects read a set of instructions on their computer screen. To make sure the participants fully understood the purpose of the experiment, they took a test and were allowed to proceed only if they answered all questions correctly. At the beginning of each round, the computer assigned each player to groups of 4. First, every player had to decide whether to opt in and play the game, or to opt out and receive a fixed 12 euro payoff. If the player opted in he received a starting budget of 10 euro and was allowed to decide how much of this 10 euro to invest in a common good. This contribution could be any natural number ranging from 0 to 10. All the money contributed to the public good was multiplied by a multiplication factor 2, which is common in the literature (e.g. Fehr, 2002). After this multiplication, the money from the common good was evenly distributed over all group members that opted in. Hence, participants that opted out of the game (also called 'loners') did not contribute to the group's common good, but also did not take any returns from it either.

In the control condition, a round only consisted of participation and contribution decisions. At the end of each round, each player got to see his earnings made in the respective round. Payoffs for non-loners were calculated as $Payoff_i = (1 - cont_i) + \sum_{i=1}^{n} cont_n/n$ , where $cont_i$ is the contribution of person *i* and *n* represents the number of persons that are participating in the game. The payoff for loners was fixed at 12 euro.

In the treatment conditions, the contribution phase was followed by a second phase where participants got to see the decisions taken by their group members in the first phase of the game. This information was released so that an informed choice concerning the punishment of a certain group member could be made. Every participant in the game got to make this choice for each of his 3 group members. This meant that loners (individuals that chose to opt out of the game in the first phase) could also punish and be punished. Both this content and the information on the cost of punishment was contained in the instructions that were shown at the start of the treatment conditions. As previously mentioned, an act of punishment caused

the punishee a cost of 6 euro while the punisher paid a price of 2 or 4 euro, depending on the condition that had to be played.

### *1.2 Programming and data analysis*

We programmed our optional public goods game using Adobe Macromedia Authorware version 7.0®. Authorware is a commonly used tool for programming visually strong economical experiments. We wrote a program that allowed for real-time interactions between the participants in the 4 different rooms by writing all the information to separate files on a common network hard drive. In supplementary figure 1, we give a taste of how the flowline of our program looks like in Authorware.

All of our data analysis was performed using R i368 2.15.0. The R code is added in the appendix.

Since we were interested in the nature of the punishment behaviour, we wanted to investigate the probability that a person punishes an interactant given certain predictor variables. The response variable is discrete and takes on the value of 1 when a player punished his interactant in the respective dyadic interaction of the game, and the value of 0 when the player did not punish his interactant. In the description of the results, we will also refer to the player having to make the punishment decision as the actor. Because of the categorical (in this case, binary) nature of the outcome variable, we use the glm() function to fit a logistic regression model to estimate the contribution of factors and variables to punishment behavior (McCullagh and Nelder*,* 1989). Since logistic regression is used to predict binary outcomes, the natural logarithm of the odds (representing the ratio of the two probabilities: those of getting 1 against those of getting 0) is used to fit the model for the predictor variables through regression analysis. Because we have more than one predictor variable, we actually fit a logistic curve in a procedure known as multiple logistic regression. The probabilities and regression coefficients were obtained using maximum likelihood estimation. In contrary to a linear regression model, where an analytical solution can be found through least squares methods, maximum likelihood estimates are found through an iterative process until convergence is achieved. To test the statistical significance of the fitted regression coefficients we used Wald statistics. These are the ratios of the fitted coefficients over the squared standard errors for those coefficients; these statistics follow a Chi-square distribution on which we can test significant deviance relative to the null hypothesis, where the coefficient would be zero.

**Table 2: Name and description of variables used for statistical modelling**

| NAME OF VARIABLE | DESCRIPTION |
|---|---|
| **COST** | CONTAINS INFORMATION ABOUT THE COST OF PUNISHMENT, WHICH BECOMES 2 IN THE LOW COST CONDITION AND 4 IN THE HIGH COST CONDITION. |
| **CONDITION** | A DISCRETE VARIABLE REFLECTING THE CONDITION AND CAN TAKE ON THE VALUES OF 'CONTROL CONDITION', 'LOW COST CONDITION' AND 'HIGH COST CONDITION'. |
| **CONTRIBUTION** | INDIVIDUAL CONTRIBUTION OF THE PLAYER WHICH CAN BE ANY NATURAL NUMBER IN THE RANGE [0,10]. |
| **CONTRIBUTION_INTERACTANT** | INDIVIDUAL CONTRIBUTION OF THE PLAYER'S INTERACTANT, WHICH CAN BE ANY NATURAL NUMBER IN THE RANGE [0,10]. |
| **LONER** | A BINARY VARIABLE PROVIDING INFORMATION ON WHETHER THE PLAYER LONED (1) OR PARTICIPATED (0). |
| **LONER_INTERACTANT** | A BINARY VARIABLE PROVIDING INFORMATION ON WHETHER THE PLAYER'S INTERACTANT LONED (1) OR PARTICIPATED (0). |
| **PAYOFF** | A CONTINUOUS VARIABLE THAT CONTAINS THE INFORMATION ON THE PLAYER'S INDIVIDUAL PAYOFF. |
| **PAYOFF_GROUP** | A CONTINUOUS VARIABLE THAT CONTAINS THE INFORMATION ON THE GROUP'S PAYOFF. |
| **PUNISHMENT** | A BINARY VARIABLE THAT DENOTES WHETHER A PLAYER PUNISHES HIS INTERACTANT. IF IT TAKES ON THE VALUE OF 1, THE INTERACTANT IS PUNISHED. IF IT TAKES ON THE VALUE OF 0, THE INTERACTANT IS NOT PUNISHED. |
| **ROUND** | A DISCONTINUOUS VARIABLE DEPICTING WHAT ROUND OF THE CONDITION IS BEING PLAYED AND CAN TAKE ON THE VALUE OF ANY NATURAL NUMBER FROM 1 TO 4. |
| **SESSION** | A VARIABLE THAT STANDS FOR THE SESSION OF THE EXPERIMENT AND CAN TAKE ON THE VALUE OF ANY NATURAL NUMBER FROM 1 TO 6. |

Based on our hypotheses that we wanted to test, there were some variables and interactions that we definitely wanted to include in our model. We give an overview of the variables that we used in our models, briefly define them and denote what values they take on in a table (Table 2). We started out fitting a very comprehensive logistic regression model with PUNISHMENT as the dependent variable and SESSION, COST, CONTRIBUTION, CONTRIBUTION_INTERACTANT, LONER, LONER_INTERACTANT as independent predictor variables. We also tested for two-way interaction effects between COST and CONTRIBUTION, CONTRIBUTION and CONTRIBUTION_INTERACTANT, CONTRIBUTION and LONER_INTERACTANT, COST and LONER, LONER and CONTRIBUTION_INTERACTANT, LONER and LONER _INTERACTANT and for three-way interaction effects between COST, CONTRIBUTION and CONTRIBUTION_INTERACTANT and between COST, CONTRIBUTION and LONER _INTERACTANT.

Because we were interested in getting the most predictive and powerful model, we used a number of algorithms that search for the best model automatically given certain criteria. First, we utilized the bestglm package to select the best model according to the AIC criteria (McLeod and Xu*,* 2010). AIC stands for “Aikake information criterion” and provides a relative measure on the amount of valuable information versus the complexity of a model. This bestglm() function left us with a model with only 4 explanatory variables: COST, CONTRIBUTION, CONTRIBUTION_INTERACTANT and LONER_INTERACTANT. Although this simple model gives the best balance between predictive power and complexity, in order to test predictive variables with less outspoken effects, we expanded this model with additional variables and interactions. For this more extensive model, there were a number of terms in the model that are not significant. To find out if we would be better off using a more simple model, we then also ran a stepwise (backward, forward) analysis on this model using the Rcmdr package (Fox*,* 2004).

Over the course of carrying out our analyses, it became clear that not punishing occurred much more frequently than punishing. Because of this sample bias, known as "zero inflation", we were afraid that our model would weigh not punishing heavier than punishing, therefore underestimating coefficients in the maximum likelihood estimation procedure. To try to resolve this, we used the logistf package and the Zelig package in R to form a model that takes this zero inflated nature of our data into account (Firth*,* 1993; Ploner *et al.,* 2006; Imai *et al.,* 2009). This package fits the regression model taking into account the prior distribution of the negative and positive examples in order to fit non-biased estimates of the regression coefficients.

We also made effect plots using the effects package in R, in order to the nature of the impact that each significant term of the model had on the outcome variable (Fox, 2003).

Another thing we wanted to take into account is which reference people take for making their punishment decision. One possibility is that people punish according to some absolute measure or norm of contribution, but another option could be that people punish others according to some relative measure, for example by comparing them to the average contribution of the group. To test this, we also built a model that took group average contribution as an extra predictor variable.

In addition, we performed an analysis of how levels of cooperation (individual contribution) developed over the course of the experiment. First, we used a Wilcoxon signed rank test to see if cooperation levels differ significantly between conditions (Wilcoxon, 1945). The null hypothesis of the Wilcoxon signed rank test assumes that the distributions of x and y differ by a location shift of μ and the alternative is that they differ by a significantly different location shift. Second, we performed a Kruskal-Wallis rank test in order to test if the sequence of the conditions has an effect on the cooperation in those conditions (Kruskal and Wallis, 1952). The null hypothesis of the test is that k independent samples were drawn from identical populations and the alternative hypothesis is that the samples were drawn from populations sharing the same shape but with different central tendencies. The Kruskal-Wallis test compared the two sessions in which the control condition came first in the experiment, with the two sessions in which the control condition came second and with those two where the control came third. We carried out the same analysis for the low and high cost punishment conditions.

We also built three linear regression models: one to predict individual contribution, one for individual payoff and another for the group's payoff. For the first two models, we again employed the bestglm() function of the bestglm-package to test which variables had the best predictive power (McLeod and Xu, 2010). We investigated how the group's payoff changed over the different conditions and over the different rounds of the game. Because we wanted to see if a raise in contribution levels would imply a raise the group's payoff, the linear regression model with GROUP_PAYOFF as a dependent variable was composed of the same predictor variables as used in the model for contribution to ease this comparison, which are SESSION, CONDITION and ROUND.

# III. Results

## 1 Descriptive analysis of punishment behaviour

Throughout all the sessions of the experiment, we observed 266 acts of punishment, which correspond to approximately 11.5% of all punishment opportunities being used (266 acts of punishment out of 2304 opportunities: 96 persons, each playing 2 punishment conditions with 4 rounds per condition, where they were able to punish 3 persons each round). Out of all the participants, 70.8% punished at least once, whereas 51% punished at least twice, 21.7% punished at least 5 times and only 6.1% punished more than 7 times (Fig. 3). The number of times that players punished others followed a Poisson distribution (λ= 2.56). Poisson distributions are usually used to model count data which results from a series of independent Bernoulli trials. These distributions are defined by a single parameter, λ, which represents the population mean and variance. One participant seemed to punish unusually frequently (22 out of 24 times), whilst all others punished less than 12 times. Based on the cumulative probability function for a Poisson distribution, the chance of that participant belonging to this same distribution was only p= 8.78e-15. Based on this, we considered this individual an outlier and removed it from all our other analyses.

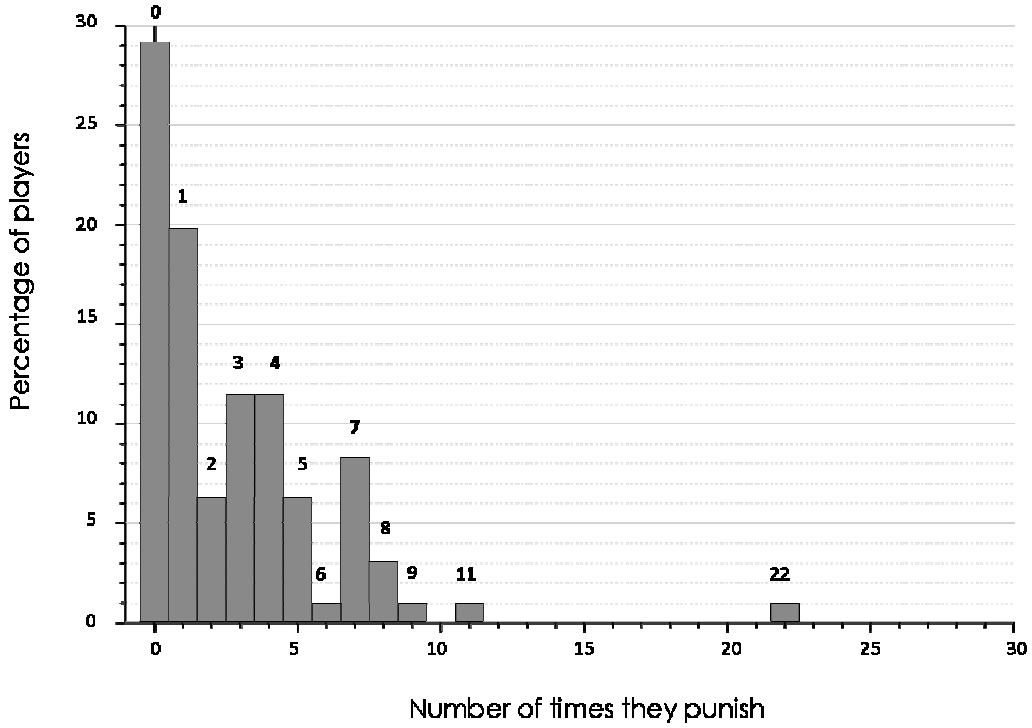



**Figure 3: Distribution of the number of times that subjects punished other group members**

## 2 Descriptive analysis of individual contribution

The average level of cooperation was higher in the treatment sessions, where punishment was possible, than in the control condition (Fig. 4). In fact, the mere threat of being punished raised the average contribution by about 3 euros. Under all conditions, the amount of money contributed to the common good decreased in the four consequent rounds. This suggests that at first, beliefs were that others would make relatively high contributions, but that when payoffs appeared to be lower than expected, people became less willing to invest in the common good. For both conditions where a punishment stage was included, punishment could not maintain the high cooperation levels of the first round. The decline in the high cost condition looks slightly steeper (Fig. 4). In the low cost condition, the cooperation level might have stabilized from the third round on. However, more rounds would be necessary to see if a steady level of cooperation would eventually be reached. To formally test for a difference in cooperation levels between different conditions, we used a Wilcoxon rank sum test with continuity correction. Contributions in the low cost condition were significantly higher than in the control condition with a p-value < 2.2e-16 and a W=27134.5. Contributions in the high cost condition were also significantly higher than in the control condition with a p-value < 2.2e-16 and a W=29264. There was no significant difference in cooperation level between the low and high cost punishment of condition (p-value=0.38 and W=59872.5).

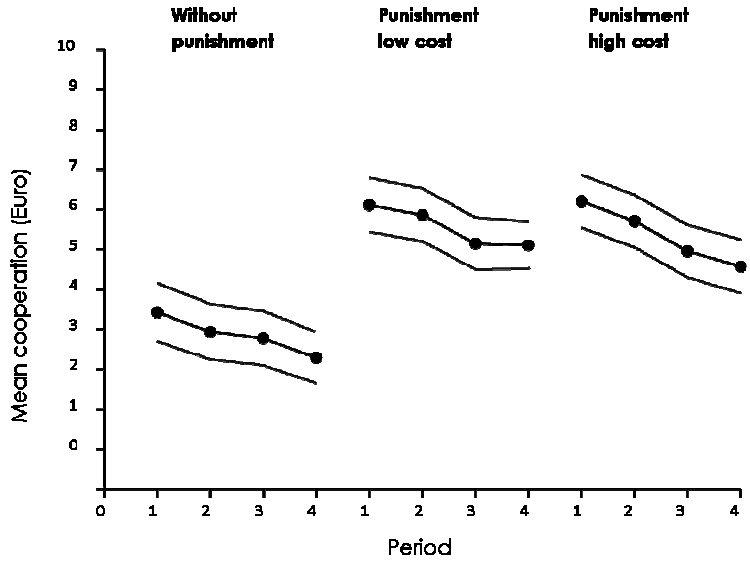



**Figure 4: Evolution of individual contribution over the 4 rounds for each condition. We conducted 6 sessions, all with the above mentioned conditions and in all possible orders.**

To find out if the sequence of treatments had an effect on the level of cooperation in each condition, we carried out a Kruskal-Wallis rank sum test. For the control condition, we reject the null hypothesis with a significance level of p=0.008 and a test statistic of 9.63. Analyses for the low cost condition (Kruskal-Wallis $\chi^2$ = 9.97 df = 2, p-value = 0.0068) and the high cost condition (Kruskal-Wallis chi-squared = 38.037, df = 2, p-value = 5.5e-09) also led us to reject the null hypothesis. Consequently, we suspect that the sequence of treatments has an effect on the level of cooperation in the respective treatments. We do think that we should be careful when it comes to interpreting these results, because all six sessions had a different order of treatments and thus we had no replicas for the different sequences. To analyze these results in more detail, however, we also performed a multiple regression analysis of individual contribution.

## 3 Multiple regression analysis of individual contribution

Our linear regression model had CONTRIBUTION as the dependent variable and SESSION, ROUND and CONDITION as predictor variables. Contributions were significantly higher in the punishment conditions compared to the control condition (Multiple linear regression, p= >2e-16, Table 3, Supplementary Fig. 3), but that the low and high cost condition did not induce a significant difference in contribution (Multiple linear regression, p=0.72, Table 4, Supplementary Fig. 3). To see whether the contribution over rounds might decrease more in the high cost than in the low cost condition we also built a model in which the two-way interaction between round and condition was included. Nevertheless, this interaction turned out not to be significant (Multiple linear regression, p=0.617, Table 5). To distinguish whether the difference in contributions across sessions is really due to the sequence of treatments or due to sampling effects we also included the interaction between session and condition in our model. Given that this interaction effect was not significant, it seems likely that differences between sessions were mostly due to sampling effects (Multiple linear regression, Supplementary Fig. 4 & 5).

**Table 3: The linear regression results for the model of individual contribution. 'ConditionH' is short for the high cost of punishment condition and 'conditionL' for the low cost of punishment condition. Both session and condition are considered factors in the model. ConditionH and conditionL are both being compared to the control condition. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | 4.621710 | 0.26335 | 17.550 | < 2e-16 | *** |
| sessie2 | -3.238840 | 0.22051 | -14.688 | < 2e-16 | *** |
| sessie3 | -0.296880 | 0.22051 | -1.346 | 0.178 | |
| sessie4 | -1.352680 | 0.22051 | -6.134 | 9.83e-10 | *** |
| sessie5 | -1.428570 | 0.22051 | -6.479 | 1.10e-10 | *** |
| sessie6 | -1.668450 | 0.22415 | -7.443 | 1.32e-13 | *** |
| round | -0.445110 | 0.05723 | -7.777 | 1.05e-14 | *** |
| conditionH | 2.600000 | 0.19549 | 13.300 | < 2e-16 | *** |
| conditionL | 2.650000 | 0.19549 | 13.556 | < 2e-16 | *** |

**Table 4: The linear regression results for the factor CONDITION in the model of contribution where the control condition and high cost condition are being compared to the low cost condition. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| conditionC | -2.65000 | 0.19549 | -13.556 | 2e-16 | *** |
| conditionH | -0.05000 | 0.13823 | -0.362 | 0.718 | |

**Table 5: The linear regression results for the two way interaction between predictor variable ROUND and factor CONDITION in the model of contribution. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| round:conditionC | 0.08737 | 0.17490 | 0.500 | 0.617 | |
| round:conditionH | -0.05579 | 0.12367 | -0.451 | 0.652 | |

## 4 Multiple logistic regression analysis of punishment behaviour

We ran a stepwise (backward, forward) analysis on our comprehensive, initial model with PUNISHMENT as the dependent variable (output values for this model are given in Table 6). Three terms were exluded from this initial model: SESSION, the two-way interaction COST and LONER and the three-way interaction COST, CONTRIBUTION and LONER_INTERACTION. The new logistic regression model that we obtained included COST, CONTRIBUTION, CONTRIBUTION_INTERACTANT, LONER, LONER_INTERACTANT as independent predictor variables and two-way interaction effects between COST and CONTRIBUTION, CONTRIBUTION and CONTRIBUTION_INTERACTANT, CONTRIBUTION and LONER_INTERACTANT, LONER and CONTRIBUTION_INTERACTANT, LONER and LONER _INTERACTANT, as well as the three-way interaction between COST,CONTRIBUTION and CONTRIBUTION_INTERACTANT. This new model's output values are shown in Table 7.

The resulting estimates from the logistf and the zelig model search function (Table 8) were very similar to those of the logistic regression model computed with the glm() function. In fact, the only difference is that the interaction between LONER and CONTRIBUTION_INTERACTANT was almost significant (p=0.057) for the zelig model, while in the normal glm model, it actually was significant (p=0.045) (Table 7). This could be explained by the fact that sampling biases did not have any significant impact on coefficient estimates while the estimation of prior distributions for both sample groups reduced the available degrees of freedom, which resulted in a very slight drop in statistical power.

**Table 6: Logistic regression results for the model of punishment**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | -2,854750 | 0,439098 | -6,501 | 7,96e-11 | *** |
| session2 | 0,462981 | 0,272557 | 1,699 | 0,08938 | . |
| session3 | -0,075751 | 0,274021 | -0,276 | 0,78221 | |
| session4 | -0,269674 | 0,275092 | -0,980 | 0,32694 | |
| session5 | 0,115515 | 0,258810 | 0,446 | 0,65536 | |
| session6 | 0,257002 | 0,263995 | 0,974 | 0,33030 | |
| cost4 | -1,135080 | 0,381953 | -2,972 | 0,00296 | ** |
| contribution | 0,397514 | 0,056477 | 7,039 | 1,94e-12 | *** |
| contribution_interactant | -0,000121 | 0,064954 | -0,002 | 0,99851 | |
| loner1 | 1,287091 | 0,627653 | 2,051 | 0,04030 | * |
| loner_interactant1 | 1,000913 | 0,538428 | 1,859 | 0,06303 | . |
| contribution:contribution_interactant | -0,053648 | 0,010428 | -5,145 | 2,680000e-07 | *** |
| contribution:loner_interactant1 | -0,398835 | 0,085033 | -4,690 | 2,73e-06 | *** |
| cost4:contribution | -0,004870 | 0,059436 | -0,082 | 0,93470 | |
| contribution_interactant:loner1 | -0,238800 | 0,122782 | -1,945 | 0,05179 | . |
| loner1:loner_interactant1 | -2,309832 | 0,928520 | -2,488 | 0,01286 | * |
| cost4:loner1 | -0,242659 | 0,766250 | -0,317 | 0,75148 | |
| cost4:contribution:contribution_interactant | 0,020581 | 0,008196 | 2,511 | 0,01204 | * |
| cost4:contribution:loner_interactant1 | 0,080259 | 0,073922 | 1,086 | 0,27760 | |

**Table 7: Output values for the stepwise built down logistic regression model for punishment**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | -2,643717 | 0,374331 | -7,063 | 1.64e-12 | *** |
| cost4 | -1,226934 | 0,328193 | -3,738 | 0,000185 | *** |
| contribution | 0,364701 | 0,052772 | 6,911 | 4,82e-012 | *** |
| contribution_interactant | -0,015895 | 0,063562 | -0,250 | 0,802531 | |
| loner1 | 1,401487 | 0,573066 | 2,446 | 0,014461 | * |
| loner_interactant1 | 1,166487 | 0,530942 | 2,197 | 0,028020 | * |
| contribution:contribution_interactant | -0,049646 | 0,009977 | -4,976 | 6.48e-07 | *** |
| contribution:loner_interactant1 | -0,370470 | 0,077924 | -4,754 | 1.99e-06 | *** |
| cost4:contribution | 0,031261 | 0,051834 | 0,603 | 0,546452 | |
| contribution_interactant:loner1 | -0,243151 | 0,121105 | -2,008 | 0,044667 | * |
| loner1:loner_interactant1 | -2,451810 | 0,917470 | -2,672 | 0,007532 | ** |
| cost4:contribution:contribution_interactant | 0,016513 | 0,007457 | 2,215 | 0,026793 | * |

**Table 8: Output values for the zelig model for punishment.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | -2.625285 | 0.374331 | -7.013 | 2.33e-012 | *** |
| cost4 | -1.209289 | 0.328193 | -3.685 | 0.000229 | *** |
| contribution | 0.361199 | 0.052772 | 6.844 | 7.68e-012 | *** |
| contribution_interactant | -0.014878 | 0.063562 | -0.234 | 0.814935 | |

| | | | | | |
|---|---|---|---|---|---|
| **LONER1** | 1.411675 | 0.573066 | 2.463 | 0.013764 | * |
| **LONER_INTERACTANT1** | 1.177133 | 0.530942 | 2.217 | 0.026619 | * |
| **CONTRIBUTION:CONTRIBUTION_INTERACTANT** | -0.049190 | 0.009977 | -4.931 | 8.20E-07 | *** |
| **CONTRIBUTION:LONER_INTERACTANT1** | -0.366749 | 0.077924 | -4.706 | 2.52E-06 | *** |
| **COST4:CONTRIBUTION** | 0.030619 | 0.051834 | 0.591 | 0.554715 | |
| **CONTRIBUTION_INTERACTANT:LONER1** | -0.230506 | 0.121105 | -1.903 | 0.056993 | . |
| **LONER1:LONER_INTERACTANT1** | -2.332555 | 0.917470 | -2.542 | 0.011010 | * |
| **COST4:CONTRIBUTION:CONTRIBUTION_INTERACTANT** | 0.016359 | 0.007457 | 2.194 | 0.028250 | * |

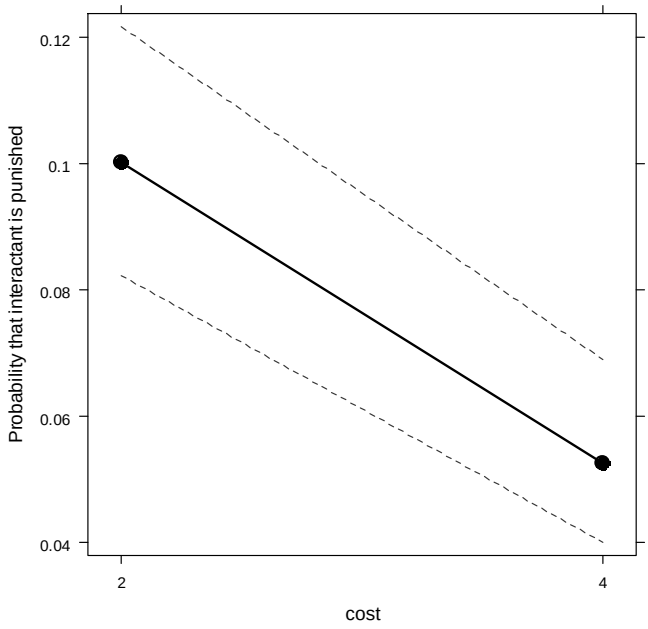



**Figure 5: Effect plot of cost**

An increase in the cost of punishment significantly decreased the probability of punishment (multiple logistic regression, Table 7, p=0.00019, Fig. 5). Contribution was positively correlated to punishment (multiple logistic regression, Table 7, p=4.82e-012, Fig. 6). In other words, cooperators punished more than defectors did. There was a significant effect of being a loner, however examination of the effect plot led us to believe loners punish almost as much as non-loners (multiple logistic

regression, Table 7, p=0.014, Fig. 7). We think that it is possible that the assumption of heteroskedasticity was not fulfilled and therefore we got seemingly conflicting results for our multiple logistic regression. Loners had less chance to be punished than the people who participated in the game (multiple logistic regression, Table 7, p=0.028, Fig. 8). The two-way interaction between CONTRIBUTION and CONTRIBUTION_INTERACTANT was highly significant (multiple logistic regression, Table 7, p=6.48e-07). When the interactant's contribution was 0, his chance of being punished was positively correlated with the actor's contribution (Fig. 9, left panel). The high contributers became less and less inclined to punish their interactant as the interactant's contribution was higher (Fig. 9, middle and right panel). Non-loners (left panel) were mainly punished by high contributors (multiple logistic regression, Table 7, p=1.99e-06, Fig. 10). The chance that a low contributor, let's say someone who contributed nothing, would punish a participant that opted into the game was very small. When the player's interactant was a loner, there was a chance that this loner would be punished if the player was a defector (left panel). However this chance was practically nonexistent if the player had cooperated (right panel). When loners punished non-loners, they punished those people whose contributions were low (multiple logistic regression, Table 7, p=0.045, Fig. 11). Loners did not punish other loners (multiple logistic regression, Table 7, p=0.0075, Fig. 12, right panel). Low contributors were punished relatively more by high contributors when the punishment cost was low (multiple logistic regression, Table 7, p=0.027, Fig. 13).

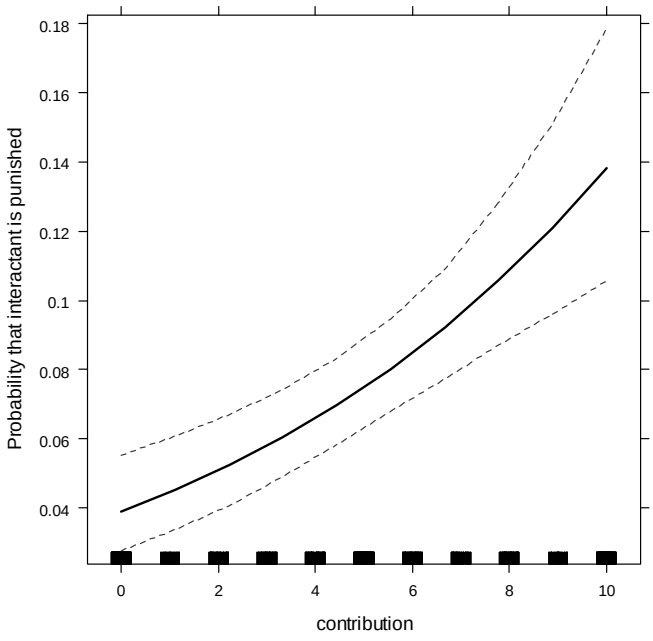



**Figure 6: Effect plot of contribution of the player who needs to make the punishment decision**

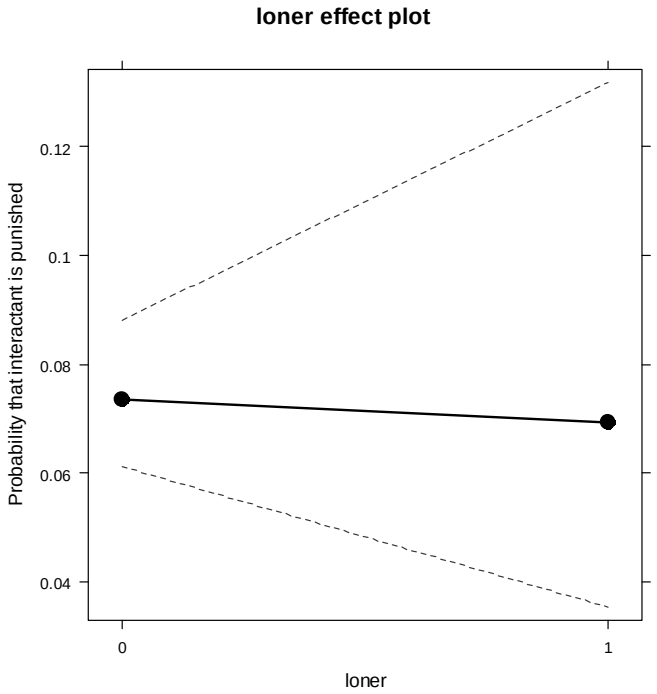



**Figure 7: Effect plot of the loner strategy of the player that is making the decision on punishment. If loner is 1, the interactant opted out of the game (x-axis).**

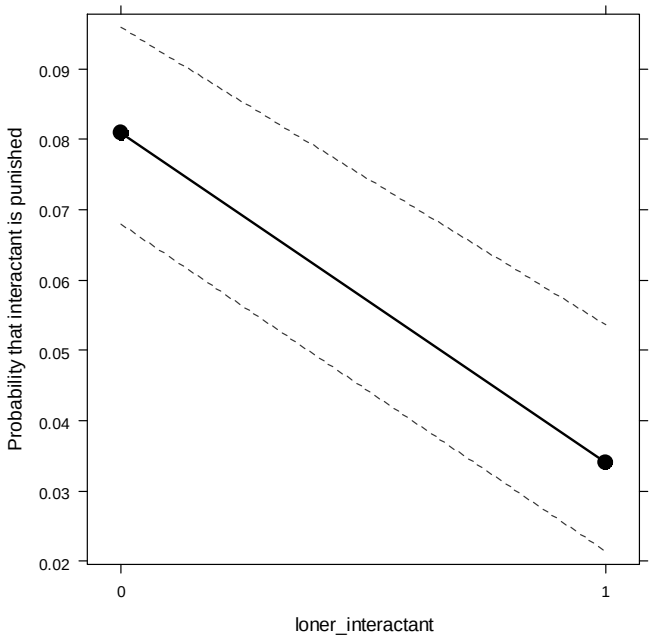



**Figure 8: The effect plot for the loner strategy of the interactant.**

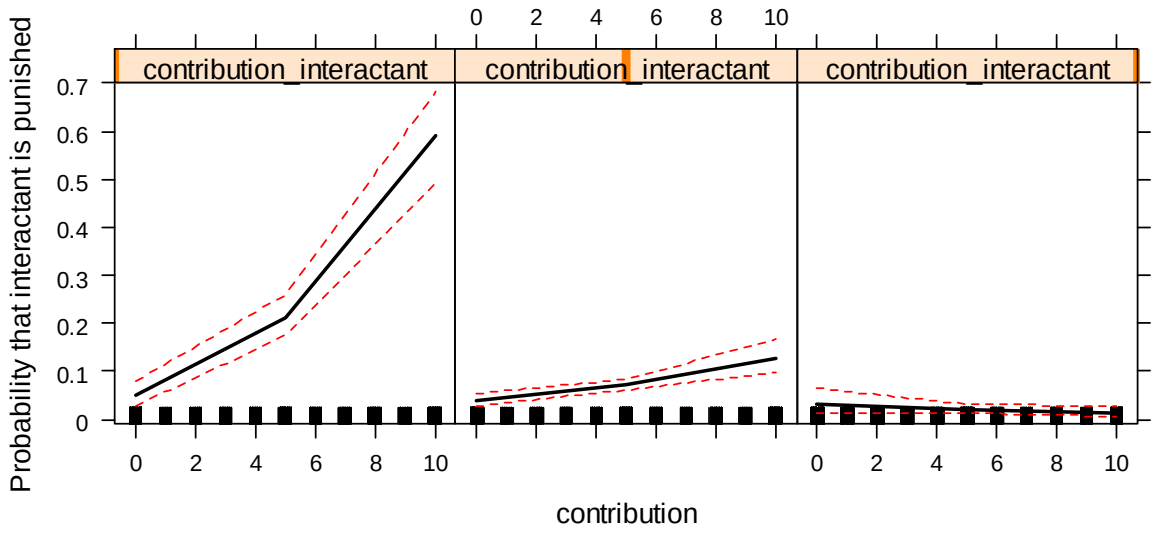



**Figure 9: Effect plot of the two-way interaction between the actor's contribution (x-axis) and the contribution of the interactant. The left panel shows the chances of punishment given your contribution and given a fixed contribution 0 euro from your interactant. In the middle panel the fixed contribution of the interactant is 5 euro and in the right panel this interactant's contribution is 10 euro.**

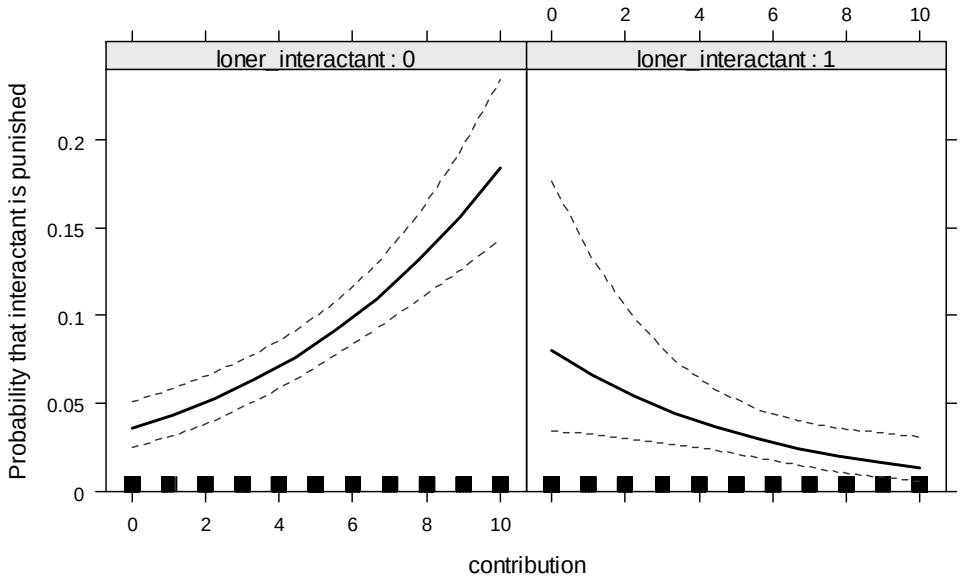



**Figure 10: Effect plot of the two-way interaction between the actor's contribution (x-axis) and the loner strategy of the interactant. The left panel shows how the chances of punishing an interactant are altered with the actor's contribution, given that your interactant opted into the game. In the right panel the same is shown for when the interactant is a loner.**

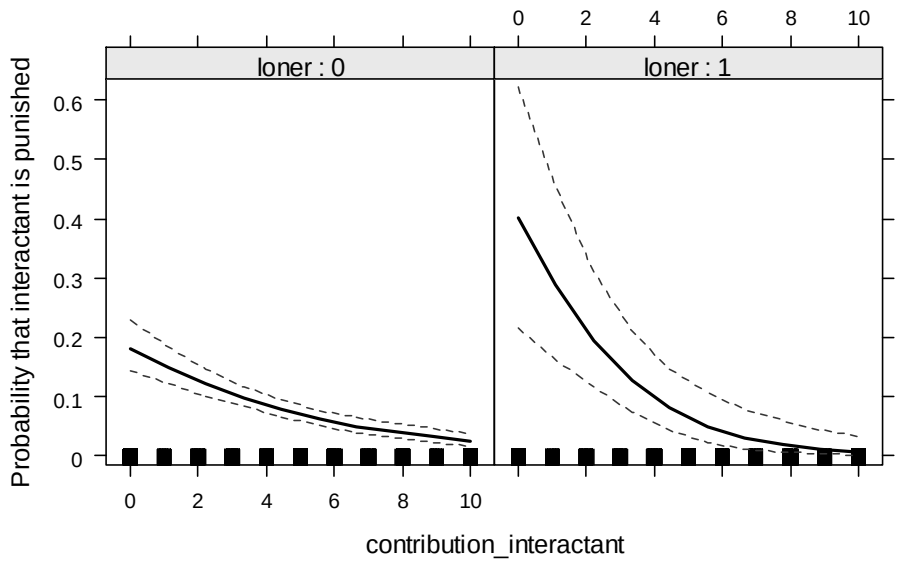



**Figure 11: Effect plot of the two-way interaction between the loner strategy of the actor and the contribution of the interactant. The panel on the right shows the chance that, given that the actor is a loner, the interactant will be punished for each possible value of that interactant's contribution.**

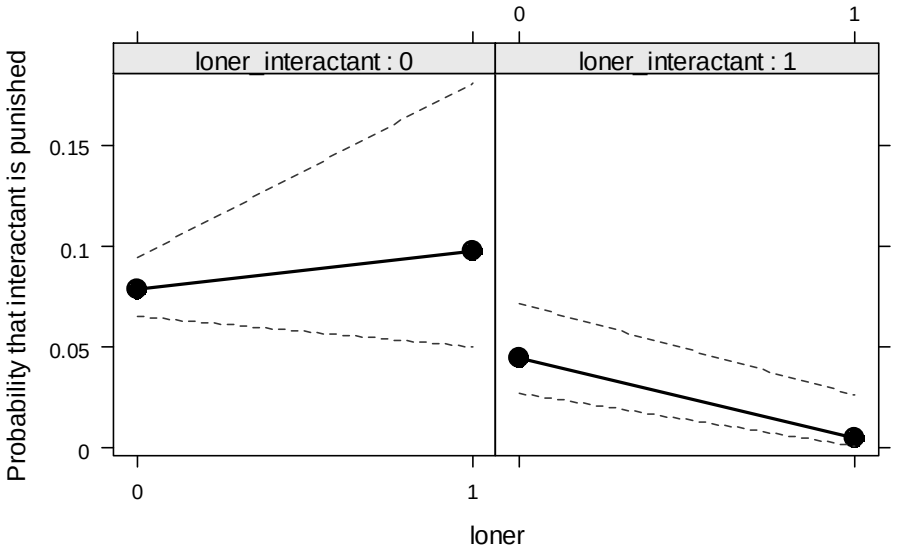



**Figure 12: Effect plot of the two-way interaction between the loner strategy of the actor and that of the interactant. The panel on the right shows the chance that, given that the interactant is a loner, the actor will punish this loner for both the strategy where the actor opts in (0 on the x-axis) and where the actor is a loner himself (1 on the x-axis).**

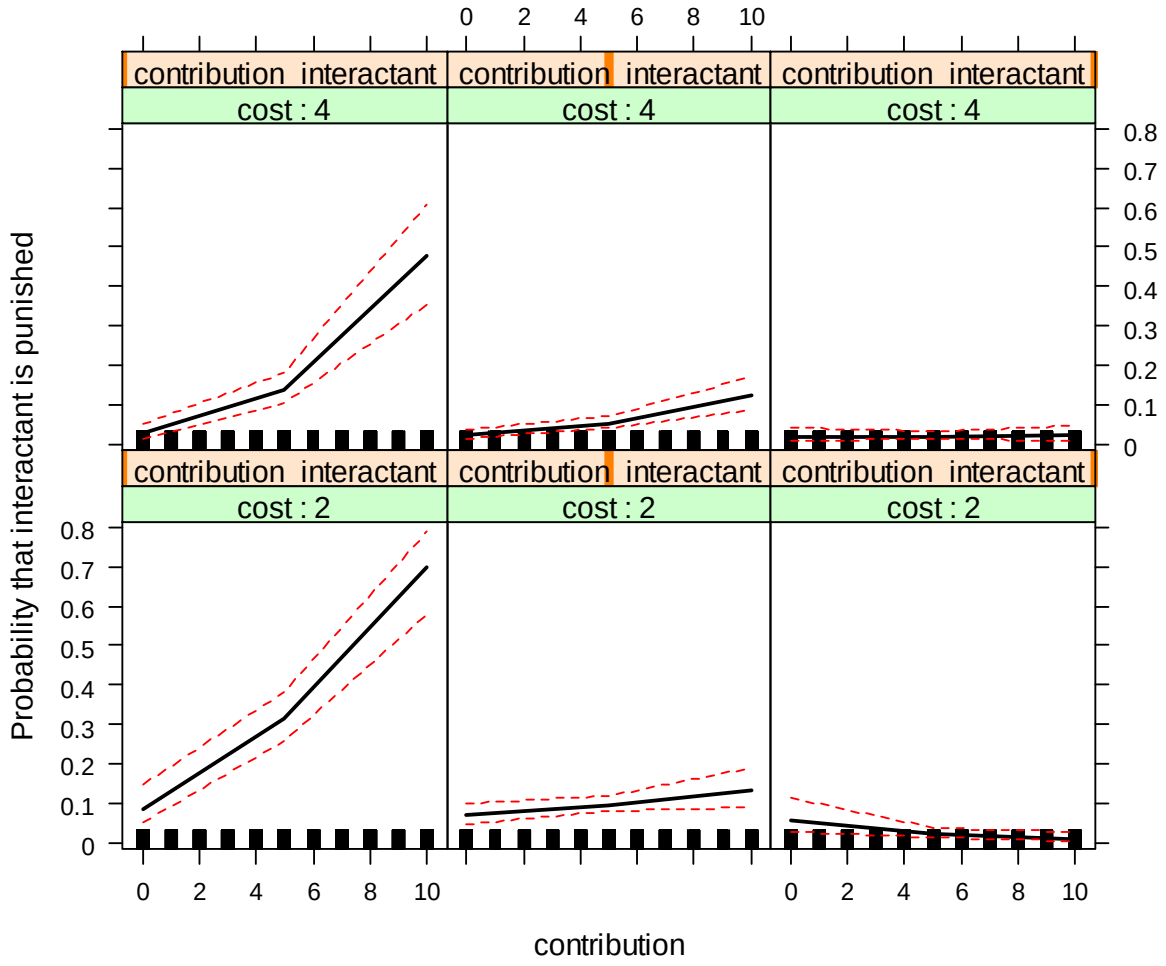



**Figure 13: The effect plot for the three-way interaction between cost, contribution of the actor and contribution of the interactant. In the upper panels, the cost is always fixed at 4 euro and in the lower panels, cost is fixed at 2 euro. In all the left hand panels the contribution of the interactant is fixed at 0 euro, in the middle panels this is fixed at 5 euro and at the right it's fixed at 10 euro.**

The average contribution of the group did not turn out to be a significant predictor variable in our multiple logistic regression model that took up CONTRIBUTION_GROUP as an extra predictor variable (Table 9, p=0.61) .

**Table 9: The logistic regression results for the punisment model including the relative measure of average group contribution**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | -2.958136 | 0.483300 | -6.121000 | 9.32e-10 | *** |
| sessie2 | 0.503671 | 0.283662 | 1.776000 | 0.07580 | . |
| sessie3 | -0.085038 | 0.274501 | -0.310000 | 0.75672 | |
| sessie4 | -0.251391 | 0.277327 | -0.906000 | 0.36468 | |
| sessie5 | 0.134485 | 0.261509 | 0.514000 | 0.60707 | |
| sessie6 | 0.268025 | 0.264838 | 1.012000 | 0.31152 | |
| cost4 | -1.127782 | 0.382065 | -2.952000 | 0.00316 | ** |
| contribution_group | 0.032081 | 0.062506 | 0.513000 | 0.60777 | |
| contribution | 0.390199 | 0.058181 | 6.707000 | 1.99e-11 | *** |
| contribution_interactant | -0.007939 | 0.066700 | -0.119000 | 0.90526 | |
| loner1 | 1.265254 | 0.629194 | 2.011000 | 0.04433 | * |
| loner_interactant1 | 0.969661 | 0.542057 | 1.789000 | 0.07364 | . |
| contribution:contribution_interactant | -0.053765 | 0.010438 | -5.151000 | 2.60e-07 | *** |
| contribution:loner_interactant1 | -0.402854 | 0.085450 | -4.715000 | 2.42e-06 | *** |
| cost4:contribution | -0.006018 | 0.059458 | -0.101000 | 0.91938 | |
| contribution_interactant:loner1 | -0.243534 | 0.123334 | -1.975000 | 0.04831 | * |
| loner1:loner_interactant1 | -2.340208 | 0.930561 | -2.515000 | 0.01191 | * |
| cost4:loner1 | -0.258597 | 0.766628 | -0.337000 | 0.73588 | |
| cost4:contribution:contribution_interactant | 0.020650 | 0.008207 | 2.516000 | 0.01187 | * |
| cost4:contribution:loner_interactant1 | 0.080998 | 0.073974 | 1.095000 | 0.27354 | |

# 5 Multiple linear regression analysis of individual payoffs

The model we got as a result of the bestglm() function for predicting a player's individual payoff contained the predictor variables SESSION, CONTRIBUTION, CONTRIBUTION_INTERACTANT, LONER, LONER_INTERACTANT and PUNISHMENT.

The payoff calculation was based on the player's own contribution, his costly punishment behaviour and that of his group members. Also, adopting a loner strategy translated into getting a fixed payoff of 12 euro. In conclusion, when intepreting the effect plots we should take into account which effects are only logical consequences of the payoff calculation structure. Session was a significant factor in the model, because the payoff is very dependent on contribution and contributions were significantly different in different sessions (multiple linear regression, Table 10). The payoff of someone contributing all his money was approximately 0.7 euro lower than that of someone not contributing at all (multiple linear regression, Table 10, p=0.015, Fig. 14). Logically, a player's payoff increased when his interactant contributed more to the common good (multiple linear regression, Table 10, p=< 2e-16, Fig. 15). Adopting a loner strategy turned out less beneficial than opting into the game (multiple linear regression, Table 10, p=< 2e-16, Fig. 16). When the player's interactant was a loner, their payoff was on average over 2 euro higher than when their interactant did not chose to opt out (multiple logistic regression, Table 10, p=8.79e-13, Fig.17). The payoff of someone who punished his interactant was much lower (M = 4.4 euro) than the payoff of someone that did not punish (multiple linear regression, Table 10, p=< 2e-16, Fig.18).

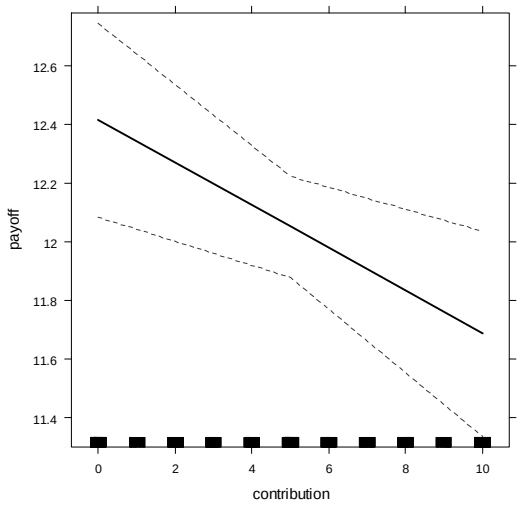



**Figure 14: The effect plot of contribution for the payoff model**

**Table 10: The logistic regression results for the payoff model.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | 11.932280 | 0.35672 | 33.450 | < 2e-16 | *** |
| sessie2 | -1.760320 | 0.33432 | -5.265 | 1.53e-07 | *** |
| sessie3 | 0.723680 | 0.30441 | 2.377 | 0.01752 | * |
| sessie4 | -0.127120 | 0.30604 | -0.415 | 0.67791 | |
| sessie5 | -0.562820 | 0.30518 | -1.844 | 0.06528 | . |
| sessie6 | -1.676690 | 0.31872 | -5.261 | 1.57e-07 | *** |
| cost4 | 0.475340 | 0.17601 | 2.701 | 0.00697 | ** |
| loner1 | -3.279340 | 0.31995 | -10.250 | < 2e-16 | *** |
| contribution | -0.072780 | 0.02992 | -2.433 | 0.01505 | * |
| loner_interactant1 | 2.323400 | 0.32316 | 7.190 | 8.79e-13 | *** |
| contribution_interactant | 0.295590 | 0.03045 | 9.707 | < 2e-16 | *** |
| interactantpunished1 | -4.661610 | 0.29741 | -15.674 | < 2e-16 | *** |

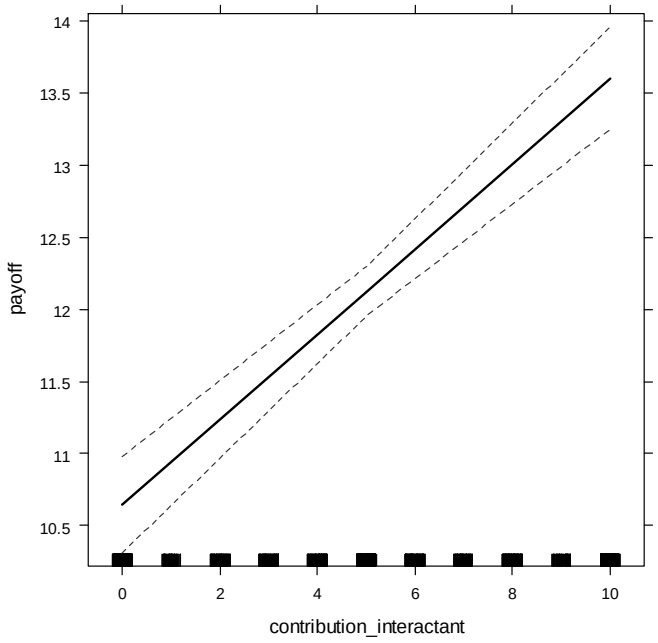



**Figure 15: The effect plot of the interactant's contribution for the payoff model**

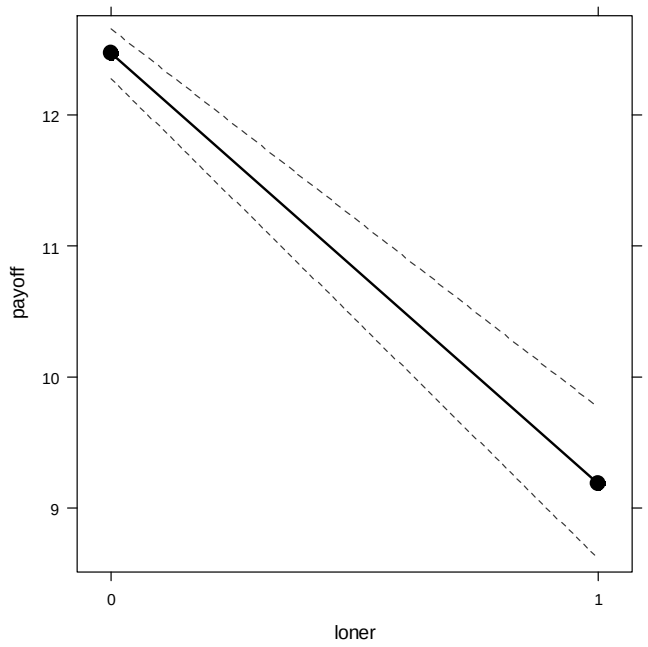



**Figure 16: The effect plot of loner for the payoff model**

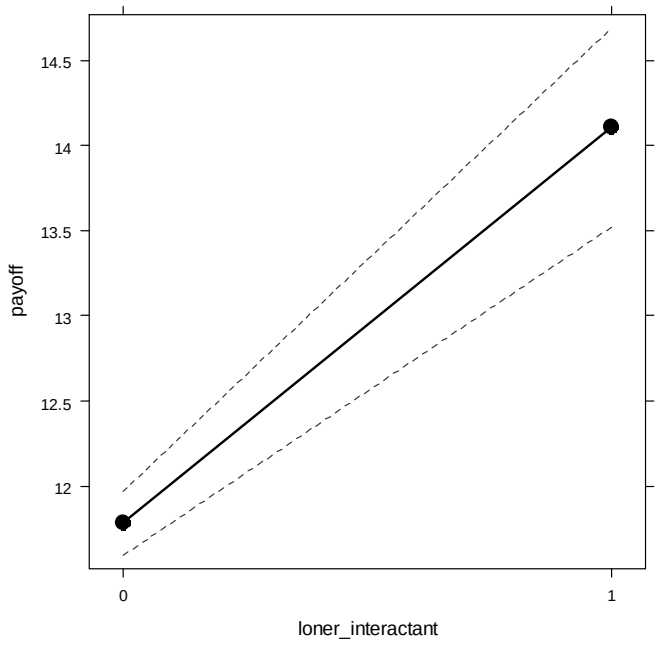



**Figure 17: The effect plot of the interactant's loner behaviour for the payoff model**

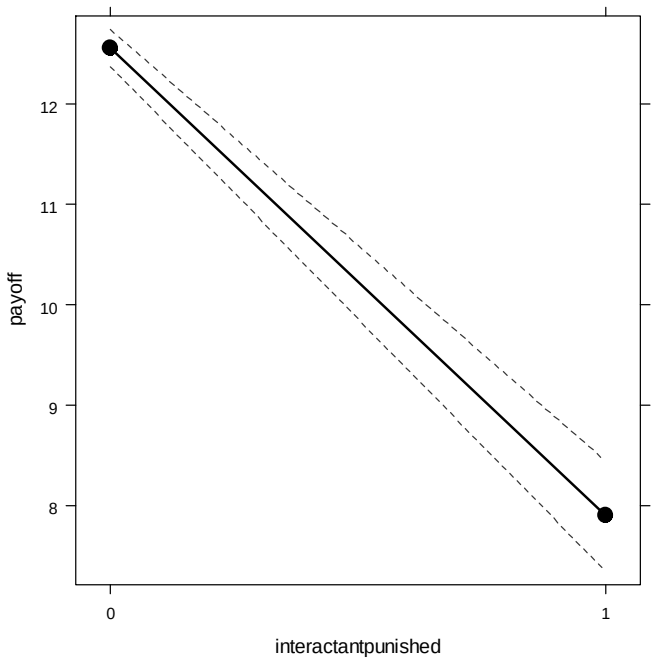



**Figure 18: The effect plot of punishment for the payoff model**

## 6 Multiple linear regression analysis of the group's total payoff

In both the low and high cost punishment condition, the group's total payoff was lower than in the control condition (Multiple linear regression, Table 11, Fig.19). In the low cost condition, the effect was truly spectacular: The drop in the group's payoff, compared to the control condition, was about 4 euro (p=5.94e-15). In the high cost condition, this drop was less than 1 euro (p=0.051).

**Tabel 11: The linear regression results for the group payoff model.**

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Signif. code |
|---|---|---|---|---|---|
| (Intercept) | 55.561800 | 0.7279 | 76.329 | < 2e-16 | *** |
| sessie2 | -10.145800 | 0.6974 | -14.548 | < 2e-16 | *** |
| sessie3 | 0.854200 | 0.6974 | 1.225 | 0.22074 | |
| sessie4 | -2.208300 | 0.6974 | -3.167 | 0.00156 | ** |
| sessie5 | -4.375000 | 0.6974 | -6.273 | 3.98e-10 | *** |
| sessie6 | -7.402800 | 0.7089 | -10.442 | < 2e-16 | *** |
| round | -0.427700 | 0.1810 | -2.363 | 0.01819 | * |

| | | | | | |
|---|---|---|---|---|---|
| **CONDITIONH** | -0.965800 | 0.4957 | -1.948 | 0.05147 | . |
| **CONDITIONL** | -3.886800 | 0.4957 | -7.841 | 5.94E-15 | *** |

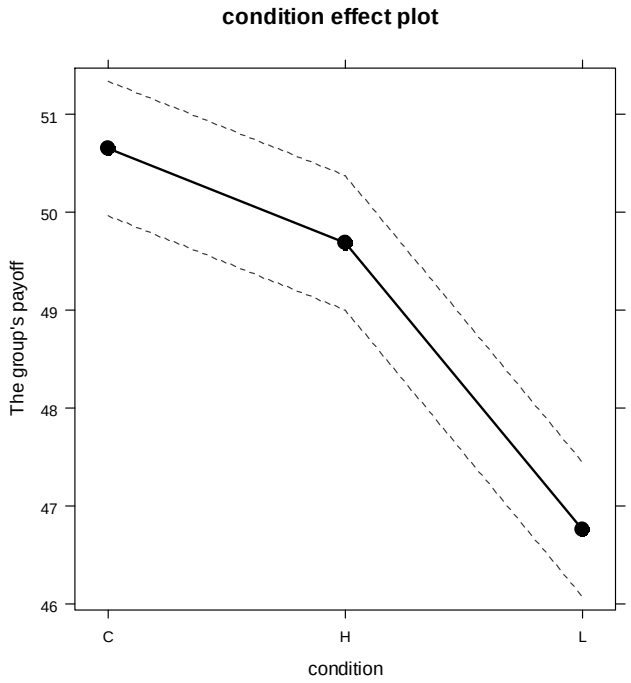



**Figure 19: Effect plot of the condition in the group's payoff**

# IV. Discussion

We conducted an experimental game in order to investigate prosocial punishment in more detail. In the literature, this has mostly been done while focusing heavily on altruistic punishment. Our study was aimed at revealing additional patterns of (prosocial) punishment, like those executed by defectors or even loners. This way, we hoped to get a better view on punishment and its effect on cooperation as a whole. Concretely, a one-shot, anonymous, optional public goods game with manipulated punishment opportunity and punishment cost was executed with 96 participants, spread over six sessions. Every participant was subjected to all following three conditions: a control condition, a low cost condition and a high cost condition.

## 1 The effect of punishment on cooperation levels

In the control condition, the average amount of money contributed to the common good decreased over the successive rounds, indicating that cooperation is not a stable outcome in a game without punishment opportunity. Logically, cooperation falls apart because there is no mechanism that prevents defectors from taking advantage of cooperators. These findings are in line with those of Fehr and Gachter (2002) and reflect the problem of common goods discussed in the literature (section II.2.3). We hypothesized that introducing the opportunity for punishment into the game would allow for cooperation to flourish. While punishment did significantly raise the individual contributions in both conditions where punishment was possible, the initially high cooperation levels did not hold over subsequent rounds. This is not what we expected. In Fehr and Gachter's (2002) study, cooperation increased over rounds and eventually stabilized, indicating that the actual execution of punishment raised cooperation levels even more than the mere threat of it. To determine whether a stable level of cooperation would be reached in our punishment conditions, we would need to include more rounds in our experimental design. However considering the experimental data that we managed to collect, we infer that the threat of punishment appeared to be very effective, yet the experience of punishment was somehow not convincing enough to maintain high cooperation levels. In order to explain the incapacity of punishment to maintain cooperation in our study, we consider two aspects of punishment: 1) its observed frequency and 2) the observed punishment strategies. The first of these two aspects is discussed here and the second one will be discussed in section IV.3.

It seems plausible that the frequency of punishment was not sufficient, when considering that during our experiment, 11.5% of all punishment opportunities were put into use, while in Fehr and Gachter's (2002) experiment, this amounted to 29.4% (1270 acts of punishment out of 4320 opportunities: 240 persons, each playing 6 rounds of a punishment condition, where they were able to punish 3 persons each round). We postulate that the cost of punishment is to blame for the relatively low occurrence of punishment observed in our experiment (section IV.2). We reckon that the observed lack of a stable cooperation level illustrates that if not enough punishers are around to split the (in our case, high) costs, punishment would fail to ensure the 'group benefit' that should be obtained through high cooperation levels (section I.3.3.2).

## 2 The effects of the cost of punishment

### *2.1 Frequency of punishment*

As expected, we observed that as the cost of punishment increased, punishment became less frequent: the augmented cost to the actor renders it harder to compensate for the same benefit to be obtained through a higher level of cooperation. Unexpectedly, even though punishment decreased in the high cost condition, cooperation levels did not significantly differ between the low and high cost conditions. A possible explanation would be that both frequencies of punishment could have led to the same pattern of cooperation over rounds because some threshold frequency of punishment exists which was not reached in our experiment. An alternative explanation is that participants were simply unaware of the real frequency of punishment and were not estimating (or sensing) the probability of being punished in function of the variable cost. As a consequence, they would just display some generic level of cooperation because they knew that they were at risk of being punished.

In section IV.1, we mentioned our suspicion that the lower frequency of punishment was responsible for the inability to maintain cooperation. This lower frequency, in its turn, could possibly be accounted for by our high cost of punishment: 2 euros in the low cost condition (20% of the endowment of 10 euro) and 4 euros (40% of the endowment) in the high cost condition. In Fehr and Gachter's (2002) experiment, the cost of punishment was not fixed; the cost of punishing a group member could be as little as 5% of the endowment (and as more money was spent on punishment, the financial loss for the punishee rose accordingly). Our decision to use high, fixed

costs was motivated by predictions, resulting from Eldakar's model (2007), that stated that selfish punishment would become obvious as the cost of punishment becomes about 40% of the endowment.

### *2.2 Profile of the punisher*

There was no shift in the punisher's profile in the different cost conditions, so the hypothesis that selfish punishers take on the task of punishing when the costs are high is not confirmed in this study. It seems like our high cost condition, instead of revealing new patterns of punishment, merely supresses punitive behaviour. This would then imply that the costs of punishment were set too high in our experiment.

We consider the previously mentioned observations concerning the cost of punishment (IV.2.1 and IV.2.2) an indication that it would be desirable to include an extra treatment condition with a lower cost of punishment (for example 1 euro or even 0.5 euro) in follow-up experiments. This suggestion for further research will also be expressed in some of the sections that follow.

## 3 Observed punishment strategies

In general, the great majority of punishment in our experiment was imposed onto defectors and we observed no antisocial punishment. Since most punishment was prosocial, the noted lack of a stable cooperation level (IV.1) cannot be accounted for by the observed punishment strategies.

We think that it is worth mentioning that the average contribution of the group did not significantly influence punishment. This led us to believe that people base their punishment decisions on some absolute idea of how much a person should contribute and not on the others' contributions relative to the group's norm. In the following subsections, we discuss the observed punishment strategies into detail.

### *3.1 Cooperators as punishers*

The most prominent punishers in our experiment were cooperators that punished defectors. This observation confirms the altruistic punishment hypothesis (Fehr and Gachter, 2002). It has been (and still is) a real challenge for theoreticians to ultimately explain altruistic punishment (section I.3). We believe that our results might be partially influenced by some constraints of the experiment itself, since subtle reputation cues have been suggested to account for altruistic behaviour

observed in anonymous experiments (Hagen and Hammerstein, 2006). For example, participants may still have felt like their reputation was at stake because of the presence of the researcher or because they knew that the experimental data was subject to careful examination. An alternative way to organize the experiment would be to let participants participate via the internet, like they do in Rand and Nowak's (2011) study. However we suspect that this method would have yet other downsides, like neither being able to control nor know the participant's environment and its associated cues.

### *3.2 Loners as punishers*

Loners also played a considerable role as punishers: they punished defectors, but not other loners or cooperators. So whereas on one hand loners could be interpreted as people who are (at least partially) asocial, they also seemed to display prosocial behaviour. These results go against the antisocial punishment hypothesis, which states that loners would be prone to punish cooperators, as models predict that cooperators could invade a group of loners (Rand and Nowak, 2011). Instead, loners enforce society's norms by engaging in prosocial punishment, even though they are not fully conforming to certain standards themselves (by not acting as cooperating participants). One possible explanation is that loners do not really lack interest in the game. They may actually want to participate, yet could be observing how the situation progresses, as a by-product of their risk-aversive nature. In particular, loners could be punishing defectors in an attempt to raise levels of cooperation before they then opt into the game. The loner's behaviour could also potentially be due to the existence of some sort of division of prosocial labour between loners who punish and cooperators who do not punish: Loners are reluctant to take part in the first common good (because they do not risk being exploited by defectors), but engage in the costs for the second common good, while cooperators who do not punish exhibit the reversed behavioural pattern. However, loners' payoffs turned out to be a lot lower than that of cooperators and point out that a division of labour is highly unlikely.

### *3.3 Defectors as punishers*

Albeit fictional scenarios and mathematical models predicted that defectors would punish other defectors (Eldakar *et al.,* 2007), we did not detect this type of selfish punishment in our experiment. This also implies that, for now, we have no reason to

believe that the evolution of altruistic punishment was facilitated by selfish punishment strategies (Nakamaru and Iwasa*,* 2006).

Defectors also did not punish cooperators, however they did punish loners to a limited degree. We do not believe this tendency to be due to a general, anti-everything mentality of defectors because we expect a more random pattern of punishment if defectors are just overall spiteful participants. However, the fact that loners are specifically targeted is consistent with predictions that arose from theoretical models from Rand and Nowak's study (2011). The predictions denoted that it would be an evolutionary stable strategy for defectors to punish loners, because loners could possibly outcompete a group of defectors. Besides, it is possible that defectors have some tendency to punish loners in order to guarantee a large enough group to exploit.

## 4 The relative success of different strategies

The payoff of a full cooperator was lower than that of a defector, however this difference in payoff was relatively small. This could be accounted for by the fact that defectors were the the most common punishees and that this effectively diminished the expected difference in payoffs between cooperators and defectors. However, defecting came out as the winning strategy, most likely because not enough punishment was carried out (section IV.1). This difference in payoffs again indicates why cooperation could not withstand over subsequent rounds under these conditions. Adopting a loner strategy turned out to be the least beneficial strategic choice of all. Although a loner received a fixed payoff of 12 euro for opting out, his average payoff was much lower because loners could also punish and be punished. We believe that researching why people adopt a loner strategy is an important task for future research. Investing whether there are game circumstances (for example, games with a different cost of punishment or a distinct fixed payoff for loners) in which it would be beneficial to display risk-aversive strategies could be a valid starting point to acclomplish that task. Additionally, analyzing how other factors, like average contribution in the game, can influence the likelihood of loning would permit a deeper insight into the profile a loner.

A punisher's payoff was much lower than the payoff of someone that didn't punish, which confirms the expectation that winners don't punish (Dreber *et al.,* 2008). On a different note, we noticed two curious things about the results that we otained. First, the difference in payoff was larger than the highest possible cost (4 euro) of

punishment. A possible cause for this result is that people could punish up to 3 interactants per round. Yet we have reason to believe that this situation was rather exceptional since the frequency of punishment in our experiment was not that high. Secondly, the session (which is interdependent with contribution) turned out to be a significant variable, even though we controlled for contribution in our model. These findings led us to conclude that there was some underlying variation in the experiment that our model did not account for.

We believe that it is a constriction of our analysis that payoff was defined at the level of the round in this particular linear regression model of individual payoff. People might display behavioural patterns that would only become clear over the sequence of rounds, since the strategy they choose in one round could very well be influenced by their experiences in the previous rounds. It would be interesting to construct a model with a new dependent variable for individual payoff that holds information on the average payoff of the whole game, and with a new predictor variable for punishment that tells us something about the total number of times that a person punished during the experiment. That way, we should be able to distinguish wether the most succesfull strategy is to punish a lot, punish a little or not punish at all.

# 5 The effect of punishment on the success of the group

Surprisingly, the group's total payoff was significantly lower when there was the possibility for punishment than when there was not. So although punishment managed to effectively bring about (temporarily) higher cooperation levels, the group was not able to reap any payoff benefits. This contradicts the idea that group selection would lead to the selection of altruistic punishment (sections I.3.1 and I.5.3.2), since punishment seems to cost the group more than it yields. Off course, we must not forget that in this game, money is the employed approximation for fitness and this does not allow for us to test how, for example, cooperation would lead to better abilities for warfare and thus produce an advantage.

Remarkably, the group's total payoff suffered mostly in the low cost condition. Cooperation levels in both the high and low cost condition were approximately the same, yet much less punishment acts took place in the high cost condition. It is very likely that the group's payoff benefited from the fact that lesser punishment acts led to lesser total costs. If punishment is much more common in lower cost conditions, one can wonder whether this would create a net benefit for the group through

increased cooperation, or a net loss though the costs of frequent punishment. This provides another argument to investigate games with a wider range of cost conditions.

# 6 Concluding remarks

Our optional public goods game with an opportunity for punishment provided evidence for prosocial punishment, but not for antisocial punishment. Our experiment did not show that punishment could maintain stable cooperation levels and the payoff of the group even seemed to suffer from the presence of punishment acts. Off course, this does not imply that punishment is not a potent mechanism to promote the evolution of cooperation. Since we think that the high costs of punishment in our study influenced many of our results, this should be taken into account in the design of further experiments.

Altruistic punishment was the most documented punishment strategy in our experiment. The most intriguing result, however, is that we found a new kind of prosocial punisher. This punisher is not a defector (also called selfish punisher), but a loner who punishes defectors. We find this puzzling for several reasons. For one, the loner chooses to enforce prosocial norms of the game, but on the other hand is at least partially reluctant to participate in the game. In particular, it is unclear how evolution would turn out a loner-punisher strategy, since loning and punishing yielded the lowest payoffs. In conclusion, we are left with more questions than answers and we believe that future research should focus on finding explanations for a loner-punisher strategy, while remaining alert for the possible interactions between the different strategies.

# V. Summary

Modern human society is characterized by an exceptional amount of cooperative behavior, such as the established support systems for the weak members of society and the fact that people are often gladly willing to help even strangers. This ‘hypersociality’ has long puzzled evolutionary biologists, psychologists and game theorists. Cooperation poses a public goods problem, yet still thrives under circumstances where established evolutionary theories, such as kin selection, direct reciprocity, indirect reciprocity and costly signaling, do not seem to suffice as an explanation. However it has been suggested that punishment of defectors, also named prosocial punishment, can maintain cooperation even in anonymous, one-shot interactions. This, in turn, raises the question of how punishment would be evolutionary stable, since it creates a second-order public goods problem. In the literature, the focus of the research on prosocial punishment has mainly been restricted to altruistic punishment, which is the punishment of defectors by cooperators. To this very day, there is no consensus about the ultimate explanation of altruistic punishment. In this study we investigated whether other punishment strategies are present, in order to obtain a more comprehensive understanding of the ways in which distinct punishment strategies could together affect cooperation levels. To do this, we conducted a one-shot, anonymous, optional public goods game with a manipulated punishment opportunity and cost of punishment. Our study revealed that although altruistic punishers were common, they were not the only ones participating in prosocial punishment: not defectors (selfish punishers), but loners engaged in a substantial share of this kind of enforcement. Despite the punishment of defectors that was observed, cooperation was not maintained in our experiment. Further investigation will clarify how cooperation, different punishment strategies and factors such as the cost of punishment are correlated.

# VI. Samenvatting

De moderne samenleving van de mens wordt gekarakteriseerd door een uitzonderlijke hoeveelheid coöperatie, denk maar aan de gevestigde systemen voor het ondersteunen van de zwakke leden van de samenleving, of aan het feit dat mensen vaak met plezier bereid zijn om iemand hulp te bieden, zelfs als die iemand een volslagen vreemde is. Deze 'hypersocialiteit' heeft reeds lang vele onderzoekers in de ban gehouden, vermits coöperatie een probleem van gedeelde goederen creëert en er toch nog in slaagt om te zegevieren onder omstandigheden waar gevestigde evolutionaire theorieën, zoals kin selectie, theorieën van directe en indirecte reciprociteit en 'costly signalling', niet lijken te volstaan als verklaring. Het is echter geopperd dat het straffen van defectoren, ook prosociaal strafgedrag genaamd, het behoud van coöperatie kan verzekeren zelfs in anonieme, eenmalige interacties. Op zijn beurt dringt zich de vraag op of straffen evolutionair stabiel kan zijn, aangezien dit strafgedrag ons confronteert met een tweede-orde probleem van gedeelde goederen. In de literatuur wordt de focus van het onderzoek naar prosociaal strafgedrag vooral gevestigd op altruïstisch strafgedrag, wat gedefinieerd is als het straffen van defectoren door coöperatoren. Tot de dag van vandaag is er geen consensus over de verklaring voor dit altruïstisch strafgedrag. In deze studie hebben we onderzocht of er andere strafstrategieën aanwezig zijn. We hopen hiermee bij te dragen aan een beter begrip van de manier waarop verschillende strafstrategieën samenspelen in het bepalen van het niveau van coöperatie. Hiertoe hebben we een 'one-shot', anoniem, optioneel gedeelde goederen spel met gemanipuleerde optie tot straffen en kost van straffen uitgevoerd. Onze studie toont aan dat altruïstische straffers het meest voorkomend zijn, maar ook dat zij zeker niet als enigen deelnemen aan prosociaal strafgedrag: niet defectoren (zelfzuchtige straffers), maar loners nemen een substantieel deel van dit strafgedrag op zich. Ondanks het waargenomen strafgedrag blijkt coöperatie niet stabiel in onze studie. Verder onderzoek moet uitwijzen hoe coöperatie, verschillende strafstrategieën en andere factoren, zoals de kost van straffen, gecorreleerd zijn.

# VII. References

Alexander, R. D. (1974). "The evolution of social behavior." *Annual review of ecology and systematics 5*: 325-383.

Alexander, R. D. (1985). "A biological interpretation of moral systems." *Zygon® 20,* (1): 3-20.

Axelrod, R. and W. D. Hamilton (1981). "The evolution of cooperation." *Science 211,* (4489): 1390-1396.

Bernhard, H., U. Fischbacher, et al. (2006). "Parochial altruism in humans." *Nature 442,* (7105): 912-915.

Bowlby, J. (1969). Attachment New York: Basic Books.

Bowles, S. and H. Gintis (2003). "Origins of human cooperation." *Genetic and cultural evolution of cooperation*: 429-443.

Boyd, R., H. Gintis, et al. (2003). "The evolution of altruistic punishment." *Proceedings of the National Academy of Sciences 100,* (6): 3531.

Boyd, R. and P. J. Richerson (1988). Culture and the evolutionary process, University of Chicago Press.

Boyd, R. and P. J. Richerson (1989). "The evolution of indirect reciprocity." *Social Networks 11,* (3): 213-236.

Boyd, R. and P. J. Richerson (1992). "Punishment allows the evolution of cooperation (or anything else) in sizable groups." *Ethol. Sociobiol. 13*: 171-195.

Boyd, R. and P. J. Richerson (2009). "Culture and the evolution of human cooperation." *Philosophical Transactions of the Royal Society B: Biological Sciences 364,* (1533): 3281-3288.

Boyd, R., P. J. Richerson, et al. (2011). "Rapid cultural adaptation can facilitate the evolution of large-scale cooperation." *Behavioral ecology and sociobiology*: 1-14.

Camerer, C. F. (2004). "Behavioral game theory: Predicting human behavior in strategic situations." *Advances in behavioral economics*: 374-392.

Camerer, C. F. and E. Fehr (2004). "Measuring social norms and preferences using experimental games: A guide for social scientists." *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*: 55-95.

Cavalli-Sforza, L. L. and M. W. Feldman (1981). Cultural transmission and evolution: A quantitative approach, Princeton Univ Pr.

Clutton-Brock, T. H. and G. A. Parker (1995). "Punishment in animal societies." *Nature 373*: 209-216.

Colwell, R. K. (1981). "Group selection is implicated in the evolution of female-biased sex ratios."

Crawford, V. P. (1997). "Theory and experiment in the analysis of strategic interaction." *Econometric Society Monographs 26*: 206-242.

Daly, M. and M. I. Wilson (2004). "Human evolutionary psychology and animal behaviour." *Animal Behaviour 57,* (3): 509-519.

Dawkins, R. (1983). The extended phenotype: The gene as the unit of selection. 1982 edition, Freeman.

de Quervain, D. J. F., U. Fischbacher, et al. (2004). "The neural basis of altruistic punishment." *Science 305,* (5688): 1254-1258.

Dreber, A., D. G. Rand, et al. (2008). "Winners don't punish." *Nature 452,* (7185): 348-351.

Eldakar, O. T., D. L. Farrell, et al. (2007). "Selfish punishment: altruism can be maintained by competition among cheaters." *Journal of theoretical biology 249,* (2): 198-205.

Eldakar, O. T. and D. S. Wilson (2008). "Selfishness as second-order altruism." *Proceedings of the National Academy of Sciences 105,* (19): 6982.

Fehr, E. and U. Fischbacher (2003). "The nature of human altruism." *Nature 425,* (6960): 785-791.

Fehr, E., U. Fischbacher, et al. (2002). "Strong reciprocity, human cooperation, and the enforcement of social norms." *Human Nature-an Interdisciplinary Biosocial Perspective 13,* (1): 1-25.

Fehr, E. and S. Gachter (2002). "Altruistic punishment in humans." *Nature 415,* (6868): 137-140.

Fehr, E. and S. Gächter (2000). "Cooperation and punishment in public goods experiments." *The American Economic Review 90,* (4): 980-994.

Fehr, E. and J. Henrich (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism.

Fehr, E. and K. M. Schmidt (1999). "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics 114,* (3): 817-868.

Firth, D. (1993). "Bias reduction of maximum likelihood estimates." *Biometrika 80,* (1): 27-38.

Foley, R. (1995). "The adaptive legacy of human evolution: A search for the environment of evolutionary adaptedness." *Evolutionary Anthropology: Issues, News, and Reviews 4,* (6): 194-203.

Fowler, J. H. (2005). "Altruistic punishment and the origin of cooperation." *Proc. Natl Acad. Sci. USA 102*: 7047-7049.

Fox, J. (2003). "Effect displays in R for generalised linear models." *Journal of Statistical Software 8,* (15): 1-27.

Fox, J. (2004). "Getting Started With the R Commander: A Basic-Statistics Graphical User Interface to R." *J. Statist. Software 14,* (9): 1-42.

Fudenberg, D. and J. Tirole (1991). Game Theory, 1991, MIT Press.

Gintis, H. (2003). "The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms." *Journal of theoretical biology 220,* (4): 407-418.

Gintis, H. (2005). Moral sentiments and material interests: The foundations of cooperation in economic life, The MIT Press.

Gintis, H., J. Henrich, et al. (2008). "Strong reciprocity and the roots of human morality." *Social Justice Research 21,* (2): 241-253.

Grafen, A. (1984). "Natural selection, kin selection and group selection." *Behavioural ecology: an evolutionary approach 2*.

Güth, W., R. Schmittberger, et al. (1982). "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior & Organization 3,* (4): 367-388.

Hagen, E. H. and P. Hammerstein (2006). "Game theory and human evolution: A critique of some recent interpretations of experimental games." *Theoretical Population Biology 69,* (3): 339-348.

Haldane, J. B. S. (1955). "Population genetics." *New Biology 18,* (3451).

Hamilton, W. D. (1964). "The genetical evolution of social behavior." *J Theor Biol 7,* (152).

Hamilton, W. D. (1975). "Innate social aptitudes of man: an approach from evolutionary genetics." *Biosocial anthropology 133*: 155.

Hardin, G. (1968). "The tragedy of the commons." *New York*.

Hardin, R. (1971). "Collective action as an agreeable n-prisoners' dilemma." *Behavioral Science 16,* (5): 472-481.

Hauert, C., S. De Monte, et al. (2002). "Volunteering as red queen mechanism for cooperation in public goods games." *Science 296,* (5570): 1129-1132.

Henrich, J. and R. Boyd (2001). "Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas." *Journal of theoretical biology 208,* (1): 79-89.

Henrich, J. and N. Henrich (2007). Why humans cooperate: A cultural and evolutionary explanation, Oxford University Press.

Imai, K., G. King, et al. (2009). "Zelig: Everyone's statistical software." *R package version 3,* (5).

Jensen, K., J. Call, et al. (2007). "Chimpanzees are rational maximizers in an ultimatum game." *Science 318,* (5847): 107.

Jones, B. and R. Sibly (1978). "Testing adaptiveness of culturally determined behavior. Do bushman women maximize their reproductive success by spacing births widely and foraging seldom." *Human behavior and adaptation. Taylor and Francis: London*: 135-157.

Kruskal, W. H. and W. A. Wallis (1952). "Use of ranks in one-criterion variance analysis." *Journal of the American statistical Association*: 583-621.

Lehmann, L., F. Rousset, et al. (2007). "Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment." *American Naturalist*: 21-36.

Levitt, S. D. and J. A. List (2007). "What do laboratory experiments measuring social preferences reveal about the real world?" *The Journal of Economic Perspectives 21,* (2): 153-174.

Loewenstein, G. F., E. U. Weber, et al. (2001). "Risk as feelings." *Psychological bulletin 127,* (2): 267.

Maynard Smith, J. (1964). "Group selection and kin selection." *Nature 201*: 1145-1147.

McCullagh, P. and J. A. Nelder (1989). Generalized linear models, Chapman & Hall/CRC.

McLeod, A. and C. Xu (2010). "bestglm: Best Subset GLM." *URL http://CRAN. R-project. org/package= bestglm*.

Nakamaru, M. and Y. Iwasa (2006). "The coevolution of altruism and punishment: Role of the selfish punisher." *Journal of theoretical biology 240,* (3): 475-488.

Nash Jr, J. F. (1950). "The bargaining problem." *Econometrica: Journal of the Econometric Society*: 155-162.

Nowak, M. A. (2006). "Five rules for the evolution of cooperation." *Science 314,* (5805): 1560.

Nowak, M. A., K. M. Page, et al. (2000). "Fairness versus reason in the ultimatum game." *Science 289,* (5485): 1773-1775.

Öhman, A. and S. Mineka (2001). "Fears, phobias, and preparedness: toward an evolved module of fear and fear learning." *Psychological review 108,* (3): 483.

Oota, H., W. Settheetham-Ishida, et al. (2001). "Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence." *Nature Genetics 29,* (1): 20-21.

Osborne, M. J. and A. Rubinstein (1994). A course in game theory, The MIT press.

Ostrom, E. (2000). "Collective action and the evolution of social norms." *The Journal of Economic Perspectives 14,* (3): 137-158.

Page, K. M., M. A. Nowak, et al. (2000). "The spatial ultimatum game." *Proceedings of the Royal Society of London. Series B: Biological Sciences 267,* (1458): 2177-2182.

Pinker, S., P. Bloom, et al. (1992). "Natural language and natural selection." *The adapted mind: Evolutionary psychology and the generation of culture*: 451-494.

Ploner, M., D. Dunkler, et al. (2006). "The logistf Package."

Price, G. R. (1972). "Extension of covariance selection mathematics." *Annals of human genetics 35,* (4): 485-490.

Price, M. E. (2008). "The resurrection of group selection as a theory of human cooperation." *Social Justice Research 21,* (2): 228-240.

Rand, D. G. and M. A. Nowak (2011). "The evolution of antisocial punishment in optional public goods games." *Nature communications 2*: 434.

Rapoport, A. and A. M. Chammah (1965). Prisoner's dilemma: A study in conflict and cooperation, Univ of Michigan Pr.

Ratnieks, F. L. W. and T. Wenseleers (2008). "Altruism in insect societies and beyond: voluntary or enforced?" *Trends in ecology & evolution 23,* (1): 45-52.

Richerson, P. J. and R. Boyd (2004). Not by genes alone: How culture transformed human evolution, University of Chicago Press.

Roberts, G. (1998). "Competitive altruism: from reciprocity to the handicap principle." *Proceedings of the Royal Society of London. Series B: Biological Sciences 265,* (1394): 427-431.

Sigmund, K. (2007). "Punish or perish? Retaliation and collaboration among humans." *Trends in ecology & evolution 22,* (11): 593-600.

Silk, J. B. (2005). "The evolution of cooperation in primate groups." *Moral sentiments and material interests: On the foundations of cooperation in economic life*: 17.

Smith, E. A. and R. L. B. Bird (2000). "Turtle hunting and tombstone opening: public generosity as costly signaling." *Evolution and Human Behavior 21,* (4): 245-261.

Smith, E. A. and R. Bliege Bird (2005). "Costly signaling and cooperative behavior." *Moral sentiments and material interests: On the foundations of cooperation in economic life*: 115-148.

Smith, E. A., M. Borgerhoff Mulder, et al. (2000). "Evolutionary analyses of human behaviour: a commentary on Daly & Wilson." *ANIMAL BEHAVIOUR-LONDON-BAILLIERE TINDALL- 60,* (4): 21-26.

Smith, E. A., M. B. Mulder, et al. (2001). "Controversies in the evolutionary social sciences: A guide for the perplexed." *Trends in ecology & evolution 16,* (3): 128-135.

Smith, J. M. (1982). Evolution and the Theory of Games, Cambridge Univ Pr.

Smith, J. M. (1986). "Evolutionary game theory." *Physica D: Nonlinear Phenomena 22,* (1–3): 43-49.

Smith, J. M. and G. Price (1973). "The Logic of Animal Conflict." *Nature 246*: 15.

Thaler, R. H. (1988). "Anomalies: The ultimatum game." *The Journal of Economic Perspectives 2,* (4): 195-206.

Trivers, R. (1971). "The evolution of reciprocal altruism." *Q. Rev. Biol. 46*: 35-57.

Wenseleers, T., A. Gardner, et al. (2010). "Social evolution theory: a review of methods and approaches." *Social behaviour: genes, ecology and evolution*: 132.

West, S., A. Griffin, et al. (2007a). "Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection." *Journal of Evolutionary Biology 20,* (2): 415-432.

West, S. A., C. El Mouden, et al. (2010). "Sixteen common misconceptions about the evolution of cooperation in humans." *Evolution and Human Behavior*.

West, S. A., A. S. Griffin, et al. (2007b). "Evolutionary explanations for cooperation." *Current Biology 17,* (16): R661-R672.

Wilcoxon, F. (1945). "Individual comparisons by ranking methods." *Biometrics Bulletin 1,* (6): 80-83.

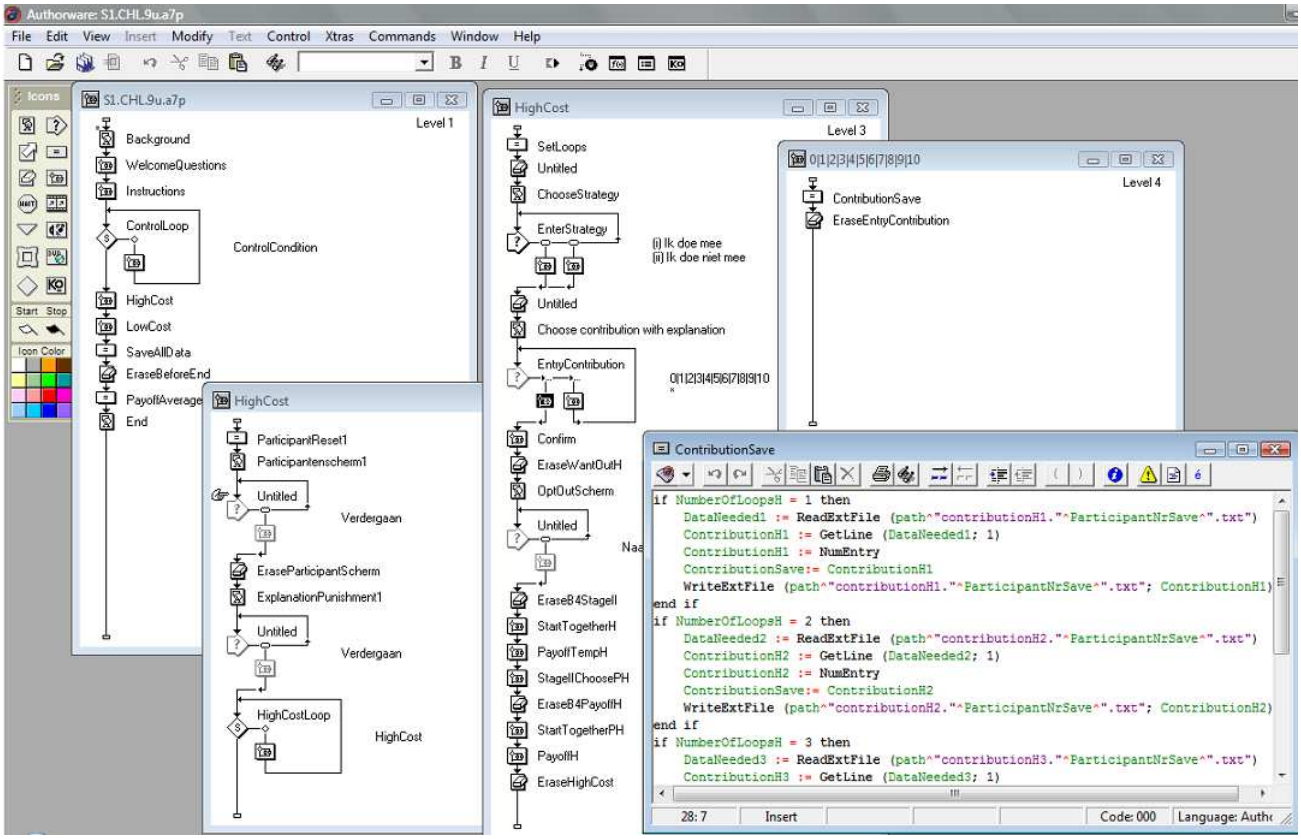
Wilson, D. S. (1997). "Introduction: Multilevel selection theory comes of age." *The American Naturalist 150,* (S1): 1-21.

Wilson, E. (1975). Sociobiology: the new synthesis, Cambridge: Harvard.

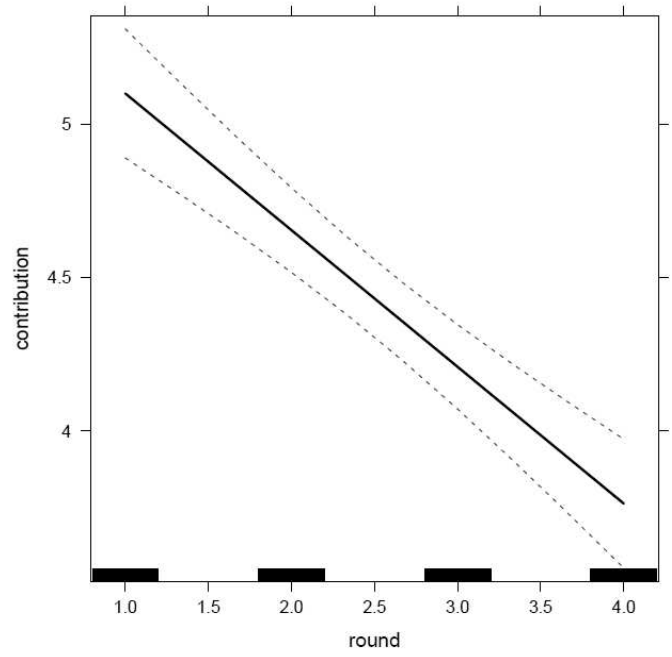
Wynne-Edwards, V. C. (1963). "Intergroup selection in the evolution of social systems."

Zahavi, A. (1975). "Mate selection--a selection for a handicap." *Journal of theoretical biology 53,* (1): 205-214.

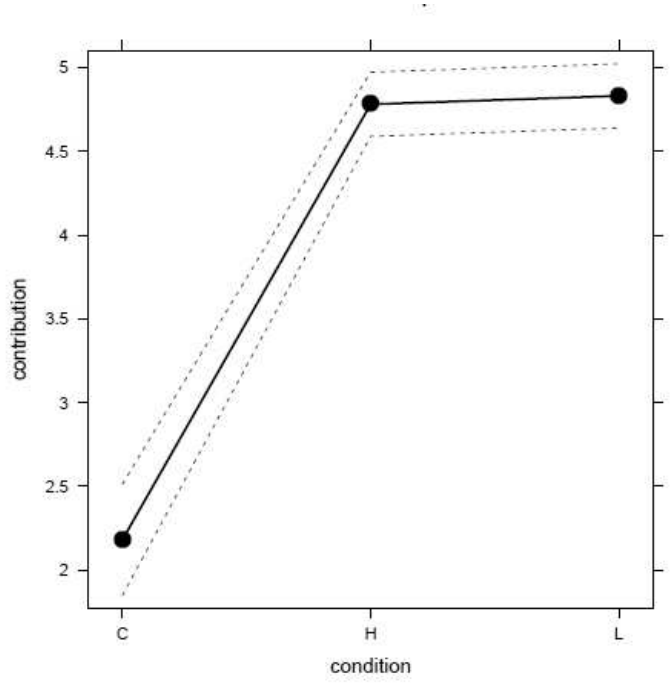
# VIII. Addendum



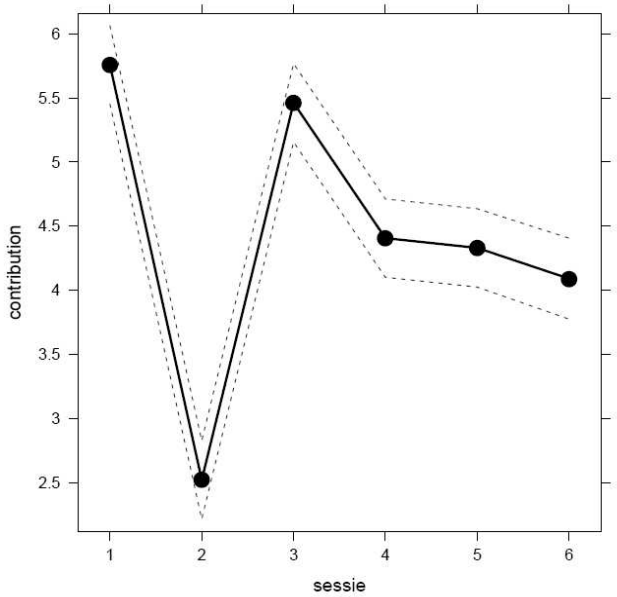

**Supplementary Figure 1: Screenshot of the program made in Authorware for the first session**



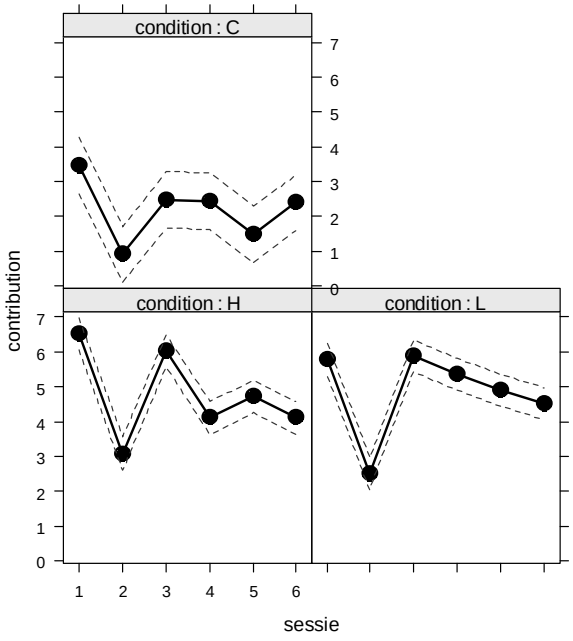

**Supplementary Figure 2: The effect plot of round for the contribution model**




**Supplementary Figure 3: The effect plot of condition for the contribution model**




**Supplementary Figure 4: The effect plot of session for the contribution model. The sequence of conditions for each session: Session 1: CHL; session 2: LCH; session 3: LHC; session 4: HCL; session 5: HLC; session 6 CLH. Conditions are labelled with abbreviations: C = control condition, L = low cost punishment condition, H = high cost punishment condition.**




**Supplementary Figure 5: The effect plot of the interaction between session and condition for the contribution model. Conditions are labelled with abbreviations: C = control condition, L = low cost punishment condition, H = high cost punishment condition.**

# R Code

```
mydata = read.table("C:\\Users\\Loren\\Dropbox\\PGG\\data analyse\\file.txt", header=TRUE, fill=TRUE)
lol <- cast(mydata2, interactantpunished ~ filenr*sessie)
counts <- t(lol)[,2]
fitdistr(counts, "Poisson")
ppois(22,lambda=2.5684211,lower.tail=FALSE)

### Wilcoxon signed rank test:
Ccontr <- mydata[mydata$condition =="C" & mydata$loner ==1,"contribution"]
Hcontr <- mydata[mydata$condition =="H" & mydata$interactant == 1 & mydata$loner ==1,"contribution"]
Lcontr <- mydata[mydata$condition =="L" & mydata$interactant == 1 & mydata$loner ==1,"contribution"]
wilcox.test(Ccontr, Lcontr)
wilcox.test(Ccontr, Hcontr)
wilcox.test(Lcontr, Hcontr)
### Kruskal-wallis test:
Ccontr1e <- mydata[(mydata$sessie == 1 | mydata$sessie == 6) & mydata$condition == "C", "contribution"]
Ccontr2e <- mydata[(mydata$sessie == 2 | mydata$sessie == 4) & mydata$condition == "C", "contribution"]
Ccontr3e <- mydata[(mydata$sessie == 3 | mydata$sessie == 5) & mydata$condition == "C", "contribution"]
Lcontr1e <- mydata[(mydata$sessie == 2 | mydata$sessie == 3) & (mydata$condition == "L" & mydata$interactant == 1), "contribution"]
Lcontr2e <- mydata[(mydata$sessie == 5 | mydata$sessie == 6) & (mydata$condition == "L" & mydata$interactant == 1), "contribution"]
Lcontr3e <- mydata[(mydata$sessie == 1 | mydata$sessie == 4) & (mydata$condition == "L" & mydata$interactant == 1), "contribution"]
Hcontr1e <- mydata[(mydata$sessie == 4 | mydata$sessie == 5) & (mydata$condition == "H" & mydata$interactant == 1),
"contribution"]
Hcontr2e <- mydata[(mydata$sessie == 1 | mydata$sessie == 3) & (mydata$condition == "H" & mydata$interactant == 1),
"contribution"]
Hcontr3e <- mydata[(mydata$sessie == 2 | mydata$sessie == 6) & (mydata$condition == "H" & mydata$interactant == 1),
"contribution"]
kruskal.test(list(Ccontr1e, Ccontr2e, Ccontr3e))
kruskal.test(list(Lcontr1e, Lcontr2e, Lcontr3e))
kruskal.test(list(Hcontr1e, Hcontr2e, Hcontr3e))

mydata$loner <- mydata$loner - 1
mydata$loner_interactant <- mydata$loner_interactant - 1
mydata2 <- mydata[mydata$condition == "L" | mydata$condition == "H",]
mydata2$sessie <- as.factor(mydata2$sessie)
mydata2$cost <- as.factor(mydata2$cost)
mydata2$loner <- as.factor(mydata2$loner)
mydata2$loner_interactant <- as.factor(mydata2$loner_interactant)
### Relogit with zelig (to take an eventual bias into account):
library(Zelig)
z.out <- zelig(interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner + loner_interactant +
contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant +
loner:loner_interactant + loner:cost + contribution:contribution_interactant:cost + contribution:loner_interactant:cost, model="relogit",
data = mydata2, tau=244/2280)
summary(z.out)
###Relogit with zelig on the built down model:
library(Zelig)
z.out <- zelig(interactantpunished ~ cost + contribution + contribution_interactant + loner + loner_interactant +
contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant +
loner:loner_interactant + contribution:contribution_interactant:cost, model="relogit", data = mydata2, tau=244/2280)
summary(z.out)
x.out <- setx(z.out)
s.out <- sim(z.out, x = x.out)
summary(s.out)
plot(s.out)

#### Logistic regression model for punishment:
interactantpunished.glm <- glm(formula = interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner +
loner_interactant + contribution:contribution_interactant + contribution:loner_interactant + contribution:cost +
```

```
loner:contribution_interactant + loner:loner_interactant + loner:cost + contribution:contribution_interactant:cost + contribution:loner_interactant:cost, family = binomial, data = mydata2)
summary(interactantpunished.glm)
### Stepwise (backward, forward) model selection:
library(Rcmdr)
stepwise(interactantpunished.glm, direction = c("backward/forward"), criterion = c("AIC"))
## Built down model (3 predictor variables less):
interactantpunished.glm <- glm(formula = interactantpunished ~ cost + contribution + contribution_interactant + loner + loner_interactant + contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant + loner:loner_interactant + contribution:contribution_interactant:cost, family = binomial, data = mydata2)
summary(interactantpunished.glm)
## Plots:
plot(effect("cost",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("contribution",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner_interactant",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("contribution:contribution_interactant",interactantpunished.glm, default.levels=3),ylab="Probability that interactant is punished",layout=c(3,1),rescale.axis=F,asp=1)
plot(effect("contribution:loner_interactant",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("contribution_interactant:loner",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner:loner_interactant",interactantpunished.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("cost:contribution:contribution_interactant ",interactantpunished.glm, default.levels=3),ylab="Probability that interactant is punished",x.var="contribution",perm.cond=c(2,1),rescale.axis=F,asp=1)
### Automated model selection:
library(bestglm)
mydata3 <- mydata2[,c('sessie','cost','loner','contribution','loner_interactant','contribution_interactant','interactantpunished')]
interactantpunished.bestglm <- bestglm(mydata3, family = binomial, IC = "AIC", method = "exhaustive")
interactantpunished.bestglm
# Plots:
interactantpunishedauto.glm <- glm(formula = interactantpunished ~ cost + contribution + loner_interactant + contribution_interactant, family = binomial, data = mydata3)
plot(effect("cost",interactantpunishedauto.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("contribution",interactantpunishedauto.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("contribution_interactant",interactantpunishedauto.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner_interactant",interactantpunishedauto.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
### Using BIC rather than AIC as the information criterion for automated search:
interactantpunished.bestglm <- bestglm(mydata3, family = binomial, IC = "BIC", method = "exhaustive")
interactantpunished.bestglm
summary(interactantpunished.bestglm)
### Stepwise forward model selection:
library(Rcmdr)
interactantpunishedforward.glm <- glm(formula = interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner + loner_interactant, family = binomial, data = mydata2)
stepwise(interactantpunishedforward.glm, direction = c("forward"), criterion = c("AIC"))

### Logistic regression with relative contributions:
mydataRel = read.table("C:\\Users\\Loren\\Dropbox\\PGG\\data analyse\\RelativeFile.txt", header=TRUE, fill=TRUE)
mydataRel$loner <- mydataRel$loner - 1
mydataRel$loner_interactant <- mydataRel$loner_interactant - 1
mydata2Rel <- mydataRel[mydataRel$condition == "L" | mydataRel$condition == "H",]
mydata2Rel$sessie <- as.factor(mydata2Rel$sessie)
mydata2Rel$cost <- as.factor(mydata2Rel$cost)
mydata2Rel$loner <- as.factor(mydata2Rel$loner)
mydata2Rel$loner_interactant <- as.factor(mydata2Rel$loner_interactant)
relative.glm <- glm(formula = interactantpunished ~ sessie + cost + relcontrib + relcontrib_interactant + loner + loner_interactant + relcontrib:relcontrib_interactant + relcontrib:loner_interactant + relcontrib:cost + loner:relcontrib_interactant + loner:loner_interactant + loner:cost + relcontrib:relcontrib_interactant:cost + relcontrib:loner_interactant:cost, family = binomial, data = mydata2Rel)
summary(relative.glm)
### Stepwise (backward, forward) model selection:
```

```
library(Rcmdr)
stepwise(relative.glm, direction = c("backward/forward"), criterion = c("AIC"))
## Reduced model (5 predictor variables less):
relativeshort.glm <- glm(formula = interactantpunished ~ sessie + cost + relcontrib + relcontrib_interactant + loner + loner_interactant + relcontrib:relcontrib_interactant + cost:relcontrib + cost:relcontrib:relcontrib_interactant, family = binomial, data = mydata2Rel)
summary(relativeshort.glm)
## Plots:
plot(effect("cost",relative.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("relcontrib",relative.glm),xlab="relative contribution",ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("relcontrib_interactant",relative.glm),xlab="relative contribution interactant",ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner",relative.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("loner_interactant",relative.glm),ylab="Probability that interactant is punished",rescale.axis=F,asp=1)
plot(effect("relcontrib:relcontrib_interactant",relative.glm, default.levels=3),xlab="relative contribution",ylab="Probability that interactant is punished",zlab="relative contribution interactant",layout=c(3,1),rescale.axis=F,asp=1)
plot(effect("cost:relcontrib:relcontrib_interactant ",relative.glm, default.levels=3),xlab="relative contribution",ylab="Probability that interactant is punished",zlab="relative contribution interactant",x.var="relcontrib",perm.cond=c(2,1),rescale.axis=F,asp=1)

#### Including the group's average contribution as a predictor variable in the model:
mydataGem = read.table("C:\\Users\\Loren\\Dropbox\\PGG\\data analyse\\file3.txt", header=TRUE, fill=TRUE)
mydataGem$loner <- mydataGem$loner - 1
mydataGem$loner_interactant <- mydataGem$loner_interactant - 1
mydata2Gem <- mydataGem[mydataGem$condition == "L" | mydataGem$condition == "H",]
mydata2Gem$sessie <- as.factor(mydata2Gem$sessie)
mydata2Gem$cost <- as.factor(mydata2Gem$cost)
mydata2Gem$loner <- as.factor(mydata2Gem$loner)
mydata2Gem$loner_interactant <- as.factor(mydata2Gem$loner_interactant)
Gem.glm <- glm(formula = interactantpunished ~ sessie + cost + GemContrib + contribution + contribution_interactant + loner + loner_interactant + contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant + loner:loner_interactant + loner:cost + contribution:contribution_interactant:cost + contribution:loner_interactant:cost, family = binomial, data = mydata2Gem)
summary(Gem.glm)
### Bestglm:
library(bestglm)
mydataBestGem <- mydata2Gem[,c('sessie','cost','loner','GemContrib','contribution','loner_interactant','contribution_interactant','interactantpunished')]
Gem.bestglm <- bestglm(mydataBestGem, family = binomial, IC = "AIC", method = "exhaustive")

### Bestmodel search for individual payoff:
mydata4 <- mydata2[,c('sessie','cost','loner','contribution','loner_interactant','contribution_interactant','interactantpunished','payoff')]
mydata4$interactantpunished <- as.factor(mydata4$interactantpunished)
payoff.bestglm <- bestglm(mydata4, family = gaussian, IC = "AIC", method = "exhaustive")
summary(payoff.bestglm)
## Plots:
payoff.glm <- glm(formula = payoff ~ sessie+ cost + loner + contribution + loner_interactant + contribution_interactant + interactantpunished, family = gaussian, data = mydata4)
plot(effect("contribution",payoff.glm, default.levels=3),rescale.axis=F,asp=1)
plot(effect("sessie",payoff.glm, default.levels=6),rescale.axis=F,asp=1)
plot(effect("interactantpunished",payoff.glm),rescale.axis=F,asp=1)
plot(effect("contribution_interactant",payoff.glm, default.levels=3),rescale.axis=F,asp=1)
plot(effect("loner",payoff.glm),rescale.axis=F,asp=1)

### Bestmodel search for individual contribution:
mydata$sessie <- as.factor(mydata$sessie)
mydata$condition <- as.factor(mydata$condition)
mydata$round <- as.numeric(mydata$round)
mydata5 <- mydata[,c('sessie','condition','round','contribution')]
contribution.bestglm <- bestglm(mydata5, family = gaussian, IC = "AIC", method = "exhaustive")
contribution.glm <- glm(formula= contribution ~ sessie + round + condition, family = gaussian, data = mydata5)
summary(contribution.glm)
```

```
mydata5$condition<-relevel(mydata5$condition, ref="L")
contributionL.glm <- glm(formula= contribution ~ sessie + round + condition, family = gaussian, data = mydata5)
summary(contributionL.glm)
## Testing interaction between condition and round:
contributioninter.glm <- glm(formula= contribution ~ sessie + round + condition + condition:round, family = gaussian, data = mydata5)
summary(contributioninter.glm)
## Plots:
plot(effect("sessie",contribution.glm),rescale.axis=F,asp=1)
plot(effect("round",contribution.glm),rescale.axis=F,asp=1)
plot(effect("condition",contribution.glm),rescale.axis=F,asp=1)

## Firth's method:
fit<-logistf(interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner + loner_interactant + contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant + loner:loner_interactant + contribution:contribution_interactant:cost, data=mydata2)
fit
summary(fit)
fit<-logistf(interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner + loner_interactant, data=mydata2)
fit2 <- logistf(formula=interactantpunished ~ sessie + cost + contribution + contribution_interactant + loner + loner_interactant + contribution:contribution_interactant + contribution:loner_interactant + contribution:cost + loner:contribution_interactant + loner:loner_interactant + loner:cost + contribution:contribution_interactant:cost, data=mydata2, family=binomial)

### Model for the group's total payoff:
payoffdata = read.table("C:\\Users\\Loren\\Dropbox\\PGG\\data analyse\\Payoffbestand.txt", header=TRUE, fill=TRUE)
payoffdata$condition <- as.factor(payoffdata$condition)
payoffdata$sessie <- as.factor(payoffdata$sessie)
payoffgroup.glm <- glm(formula = Payoff_group ~ sessie + round + condition, family = gaussian, data = payoffdata)
summary(payoffgroup.glm)
## Plots:
plot(effect("sessie",payoffgroup.glm),ylab="The group's payoff",rescale.axis=F,asp=1)
plot(effect("round",payoffgroup.glm),ylab="The group's payoff",rescale.axis=F,asp=1)
plot(effect("condition",payoffgroup.glm),ylab="The group's payoff",rescale.axis=F,asp=1)
```