Cooperation and Punishment in Humans: Exploring the Effect of Power Asymmetries and the Motivations Underpinning Punishment

By

Jonathan E Bone

A thesis submitted to University College London for the Degree of Doctor of Philosophy

Declaration of Originality

I, Jonathan Bone confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

People willingly pay to harm cheats in economic games. Although, punishment ostensibly increases cooperation levels, consensus is lacking over when punishment can increase individual or group payoffs and what motivates punishment decisions. Most previous studies have assumed that all individuals are equal. However, in reality individuals often vary in terms of power, such that some players are able to inflict a greater cost on their partner than their partner is able to reciprocate. I investigated the effect of power asymmetries on cooperation and punishment in repeated prisoner's dilemma games with punishment both where cooperation investment was binary and where cooperation investment was variable. I found that punishment did not promote cooperation from targets in any conditions.

Several studies have suggested that punishment may be motivated by disadvantageous inequality aversion. These findings raise the possibility that individuals use punishment to restore equality. However, the alternative that punishment is simply motivated by a desire for revenge and is not tailored to achieve equality, cannot be ruled out. I used a modified dictator game with punishment to disentangle these two possibilities. I found evidence that punishment was motivated by both a desire for revenge and a desire for equality.

Individuals often punish those who deviate from social norms. Why atypical behaviour is more likely to be punished than typical behaviour remains unclear. One possibility is that individuals simply dislike norm violators. Alternatively, individuals may be more likely to punish atypical behaviour because the cost of punishment generally increases with the number of individuals punished. To test these hypotheses, I used a modified public goods game with third party punishment. My results suggest that punishment of atypical behaviour might often be explained in terms of the costs to the punisher, rather than responses to norm violators.

In summary, my thesis sheds light on the conditions in which punishment is most likely to promote cooperation and on the motivations underpinning punishment decisions.

Acknowledgments

I wish to thank my supervisor Nichola Raihani for the advice and support she provided me throughout my time as her student. I have been very lucky to have a supervisor who cared a great deal about my work and always responded to my questions promptly. I would also like to thank my other collaborators Antonio Silva, Redouan Bshary, Brian Wallace and Katherine McAuliffe for their helpful suggestions and discussion. I also wish to thank Brian Wallace for allowing me to use the UCL, Department of Economics computer laboratory to run experiments and for helping me to do so.

I would like to thank CoMPLEX for the studentship which has allowed me to pursue this PhD. I would also like to thank all the members of staff at CoMPLEX for supporting me throughout. In particular, I would like to say thank you to Buzz Baum and Max Reuter for their on-going support and encouragement and to the course administrators who were always there to help with any queries I had. I wish to thank all the other students at CoMPLEX for their friendship since we all joined the MRes course, what now seems like a very long time ago!

My family have given me the opportunity to be in the position to carry out this research for which I will always be truly grateful. I would like to thank my brother James Bone and my dad Steve Bone, whose own PhD achievements gave the confidence to pursue a PhD myself. I would also like to thank my mum Janette Bone for always being there to listen to my problems and for patiently proof reading my work. I also want to express my appreciation for my girlfriend Charlotte who supported and encouraged me, even in the busy months of writing up when I had very little to give her.

Finally I want to thank the many individuals who volunteered to take part in my experiments, without their participation my research would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 The evolution of cooperative behaviour

The human tendency to cooperate with non-relatives, including complete strangers who we are unlikely to meet again in the future, exceeds that of all other species. Following Bshary & Bergmüller (2008), I define cooperation as a social interaction which results in lifetime fitness benefits to interacting individuals. Cooperation often requires individuals to pay a short-term cost in order to provide benefits for another individual (Bshary & Bergmüller 2008; West et al. 2007). In this way, cooperative behaviour can be thought of as a kind of investment (Bshary & Bergmüller 2008). Since, selection favours individuals that maximize their own fitness (Darwin 1859), the problem facing evolutionary biologists is to explain how individuals that make short-term cooperative investments are compensated in terms of lifetime fitness.

One way in which cooperative investments can be repaid is through indirect fitness benefits. Cooperative individuals can receive indirect fitness benefits by helping relatives to reproduce and in doing so, helping shared genes get passed on to the next generation (known as 'Kin Selection'; Hamilton 1964a; Hamilton 1964b; Hamilton 1975; Hamilton 1972). Nevertheless, in several species (especially humans) cooperative behaviour also exists between unrelated individuals. Under these circumstances, individuals must somehow derive direct fitness benefits from cooperative behaviours.

A traditional approach to investigating how cooperative behaviour can result in direct fitness benefits has been to use simple economic games. Of these games, perhaps the best known is the prisoner's dilemma (PD; Luce & Raïffa 1957; Rapoport & Chammah 1965). The PD describes a hypothetical situation involving two individuals who must decide whether to cooperate or cheat (commonly referred to as 'defect' in the PD setting; Luce & Raïffa 1957): if both players cooperate they gain more than if both defect. However, in a one shot game, each player does their best to defect, regardless of their partner's behaviour (Figure 1.1). Mutual defection is therefore the unique Nash equilibrium (see glossary; Nash 1951) and the only evolutionarily stable strategy (see glossary; Maynard Smith 1982) in a one-shot game (Luce & Raïffa 1957). The public goods game

(PGG; Ledyard 1995) is closely related to the PD game but is used to study cooperation in a group setting. In the PGG, players are endowed with an initial sum of money and must decide how much money to contribute to a communal pot. The communal pot is then multiplied by a factor (greater than one and less than the number of players, N), producing the 'public good'. The public good is then divided evenly amongst the players regardless of how much they contributed. In the PGG, the collective payoff is maximized when everyone contributes the maximum amount to the public good. However, individuals do best by contributing nothing and 'free-riding' on the investments made by the others in the group. Thus, if everybody adopts the 'rational' strategy of contributing nothing, no public good is produced. It has been argued that the social dilemma faced by individuals playing the PD game and the PGG is relevant to real-world cooperation problems due to their resemblance to many activities that have been important throughout our evolutionary history; for example, hunting big game, food sharing, trade, conserving common property resources, and warfare (Fehr & Gächter 2002).

		Player 1	
		Cooperate	Defect
Player 2	Cooperate	R,R	T,S
	Defect	S,T	P,P

Figure 1.1 The payoff matrix of the prisoner's dilemma game. Mutual cooperation pays each player a reward R, whereas mutual defection yields the punishment P. If a cooperator interacts with a defector, the cooperator gets the 'sucker's payoff', S (the lowest pay-off in the game) and the defector gets the 'temptation to defect', T (the highest payoff). In order to be a prisoners dilemma, the pay-offs must satisfy $T > R > P > S$.

The study of these, along with other economic games, has uncovered a variety of mechanisms by which individuals can directly benefit from cooperative behaviours, including but not limited to direct reciprocity (Axelrod 1984; Trivers 1971), indirect reciprocity (Panchanathan & Boyd 2004; Nowak & Sigmund 1998; Alexander 1987; Wedekind & Braithwaite 2002; Nowak & Sigmund 2005; Boyd & Richerson 1989; Seinen & Schram 2006; Rockenbach & Milinski 2006; Milinski et al. 2002), partner choice (Sylwester & Roberts 2010; Barclay & Willer 2007; Roberts 1998; Barclay 2004; McNamara et al. 2008; Hardy & Van Vugt 2006; Sylwester & Roberts 2013) and punishment (Gardner & West 2004a; Henrich & Boyd 2001; Boyd et al. 2003; Gintis 2000; Fehr & Gächter 2000; Fehr & Gächter 2002; Gächter et al. 2008). Trivers (1971) proposed that when individuals are likely to interact repeatedly, cooperative behaviour could evolve among non-relatives if players are conditionally cooperative. Specifically, direct reciprocity works on the principle that 'if I help you now, you will be more likely to help me in the future'. In a repeated version of the PD game ('known as an iterated prisoners dilemma'; IPD), direct reciprocity has been shown to generate stable cooperation. However, in the absence of other control mechanisms, direct reciprocity commonly breaks down in larger groups (e.g. in repeated PGG's) because it is not possible to retaliate against non-cooperators without also harming cooperative group members (Boyd & Richerson 1988).

Nevertheless, empirical work suggests that humans will often behave cooperatively even in one-shot encounters, where there is no scope for direct reciprocity to occur (e.g. Fehr et al. 2002; Fehr & Henrich 2003; McCabe et al. 2003; Henrich et al. 2005). In one-shot encounters, indirect reciprocity provides an alternative mechanism for the evolution of cooperation (Panchanathan & Boyd 2004; Nowak & Sigmund 1998; Alexander 1987; Wedekind & Braithwaite 2002; Nowak & Sigmund 2005; Boyd & Richerson 1989; Seinen & Schram 2006; Rockenbach & Milinski 2006; Milinski et al. 2002). Indirect reciprocity requires individuals to build a reputation based on their history of cooperation with others and works on the principle that 'if I help you now, then it is more likely that a third individual (i.e. a bystander) will help me in the future'. The ability to build a reputation may also promote cooperation through partner choice (Sylwester &

Roberts 2010; Barclay & Willer 2007; Roberts 1998; Barclay 2004; McNamara et al. 2008; Hardy & Van Vugt 2006; Sylwester & Roberts 2013; Noë & Hammerstein 1995). According to this theory, when individuals seek to acquire the best cooperators as partners, there will competition to be more cooperative than others in order to gain access to cooperative partners (Roberts 1998; Barclay & Willer 2007; Noë & Hammerstein 1995). Therefore, the most cooperative individuals can incur benefits from an increased access to cooperative partners. Nevertheless, these mechanisms, cannot explain why humans frequently cooperate with large groups of genetically unrelated individuals, when reputational gains are small or absent. It has been suggested that punishment may provide a solution to this problem (Gardner & West 2004a; Henrich & Boyd 2001; Boyd et al. 2003; Gintis 2000; Fehr & Gächter 2000; Fehr & Gächter 2002; Gächter et al. 2008).

1.2 Punishment and cooperation

Punishment typically involves an individual paying a cost in order to harm a peer who previously cheated (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012). Punishment can therefore be thought of as negative reciprocity. Punishment may sustain cooperation by imposing costs on cheats so that it pays for them to avoid cheating again in subsequent interactions with the punisher (Clutton-Brock & Parker 1995; McCullough et al. 2013). This increased cooperation provides a possible mechanism by which the costs associated with punishment can be recouped by punishers in two-player games. However, in larger groups, the benefits resulting from the punishment of free-riders are shared amongst the group. Therefore, it has been suggested that punishment itself represents a 'second order' public good, where individuals will be tempted to abstain from punishing (even if they contributed to the public good), whilst free-riding on the other group members' punishment investments (Yamagishi 1986; Colman 2006). This, it has been argued, will result in the invasion of punishers, destabilizing punishment and leading to reduced cooperation (Sigmund 2007; Rankin et al. 2009; Gardner & West 2004a; Panchanathan & Boyd 2004; Hauert et al. 2007; Hilbe & Sigmund 2010; Boyd et al. 2010). While some theorists have

suggested that the second order public goods problem may be solved by 'second-order punishment', the punishment of those who do not punish cheats (Axelrod 1986; Henrich 2004; Henrich & Boyd 2001), this solution only shifts the problem up a level because 'third-order punishment' would be needed to maintain second-order punishment and so on. Nevertheless, recent theoretical work has suggested that within a wide parameter space typical of experimental games, in the long term, the individual benefits arising from increased cooperation when any group member punishes will offset the costs of punishing (even in groups of larger than two; Roberts 2013). This means that in repeated games, punishment may often be in the punisher's self-interest, without the need for any higher order punishment (Roberts 2013; Bshary & Bshary 2010).

Other theoretical models have demonstrated that when individuals are able to build a reputation, punishment can evolve because carrying a reputation as someone who is prepared to invest in punishment reduces the risk that future partners will defect (Brandt et al. 2003; Hilbe & Traulsen 2012; Sigmund et al. 2001; Hilbe & Sigmund 2010; dos Santos et al. 2011; Roos et al. 2014). In addition, Raihani & Bashary (2015) argue that when competitive motives can be ruled out, a punitive reputation may convey an individual's cooperative intent. If partner choice is an option, individuals are expected to preferentially choose cooperative players as interaction partners (Sylwester & Roberts 2010; Barclay & Willer 2007; Roberts 1998; Barclay 2004; McNamara et al. 2008; Hardy & Van Vugt 2006; Sylwester & Roberts 2013; Noë & Hammerstein 1995). Therefore, punishers may benefit from being preferentially selected for interactions (Raihani & Bshary 2015). However, these mechanisms do not explain why punishment is also observed in one-shot games where there is no opportunity to build for players to build punitive reputation (e.g. Fehr & Gächter, 2002; Gächter & Herrmann, 2009; Walker & Halloran, 2004).

Several authors have suggested that cultural processes - as well as genetic - may underpin the evolution of punishment (e.g. Boehm, 1993, 1999; Gächter, Herrmann, & Thöni, 2010; Joseph Henrich et al., 2006). This has lead to the development of a cultural group selection explanation for the evolution of

punishment (Boyd et al. 2003; Gintis et al. 2003; Bowles & Gintis 2004; Boyd & Richerson 1982; Boyd & Richerson 1990; Falk et al. 2005; Fehr et al. 2002; Henrich 2004; Henrich & Boyd 2001; Lehmann et al. 2007; Boyd & Richerson 2005). According to cultural group selection, cultural traits can quickly spread through groups via social learning and when groups are in competition with each other, groups with cultural traits that are beneficial to the group, like punishment, will outcompete and replace groups without these traits. In this way, group-level benefits can offset the individual costs associated with punishment, even in one-shot anonymous encounters. It has been argued that punishment of cheating group members can evolve through cultural group selection more easily than cooperation itself because when cheating is rare the costs of punishment become low, so punishers only have a weak payoff disadvantage relative to non-punishers (Boyd et al. 2003; Henrich & Boyd 2001). Although in theory, group-level benefits could also arise through genetic group selection, maintaining high enough between-group genetic variation and strong enough between-group selection pressure for this to occur would require unrealistically low rates of migration and unrealistically high rates of group extinction (Ridley 2004). The requirements for cultural group selection however are less strict because cultural transmission of traits throughout a group is typically much faster than genetic inheritance (due in part to the fact that we can inherit genetic traits from individuals other than our parents), meaning that the required levels of between-group variation can arise (El Mouden et al. 2014; Bell et al. 2009). In addition, groups performing less well than the other groups may simply adopt the winning groups' behavioural traits rather than becoming extinct (El Mouden et al. 2014; Bell et al. 2009). Nevertheless, there is still considerable debate over the plausibility of the assumptions underpinning the cultural group selection account of the evolution of punishment (Burnham & Johnson 2005; Hagen & Hammerstein 2006; West et al. 2010). Several authors have argued that since punishment is altruistic in both cultural and genetic group selection models, it requires kin selection to spread throughout a population (Lehmann et al. 2007; Gardner & West 2004a) and therefore adds nothing useful to our understanding of how punishment can spread through populations of individuals with low relatedness (Foster et al. 2006). In

addition, these models cannot account for the initial establishment of punishment when it is initially rare (Gardner & West 2004a; Lehmann et al. 2007; Fowler 2005) or explain why individuals punish even when punishment reduces group payoffs (e.g. Fehr & Gächter, 2002).

An alternative explanation for the observation that people often punish in anonymous one--shot games is that such punishment occurs as a result of the miss-firing of psychological mechanisms which evolved (or were learnt) in the context of repeated interactions where interactions were typically observed by others (Ben-Ner & Putterman 2000; Burnham & Johnson 2005; Cosmides & Tooby 1989; Delton et al. 2011; Hagen & Hammerstein 2006; Hoffman et al. 1998; Johnson et al. 2003; Tooby et al. 2006). Consequently, these psychological mechanisms may motivate players to punish in one-shot lab settings where such behaviour is not in their best monetary interests. Nevertheless, laboratory studies using one-shot games have demonstrated that when players are given time to 'cool off' before making their punishment decision they are less likely to punish cheating partners (Grimm & Mengel 2011; Smith & Silberberg 2010; Sutter et al. 2003). These studies suggest that when players are given time to consider their decisions they are more likely to respond in a way that maximizes their payoff in their current one-shot setting rather than rely on intuitions that may maximize payoffs over repeated encounters in the real world.

Despite the myriad of studies demonstrating people's willingness to engage in punishment under laboratory conditions, real-life evidence of punishment in humans is relatively rare. In his seminal paper, Boehm (1993) surveyed the ethnographic literature revealing that punishment occurs in wide range of forms; from ridicule, gossip and verbal reproach up to social ostracism and homicide. Boehm (1993)'s review focused on small-scale egalitarian societies without centralized leadership, such societies are likely to resemble those in which humans lived for most of their evolutionary history (Sahlins 1972). However, when groups become larger and repeated interactions become infrequent, as in most modern societies, it is more difficult for individuals to recoup the costs associated with punishing cheats. Consequently, in larger populations,

decentralized punishment (see Glossary) is expected to be uncommon unless the costs associated with punishment are low (Balafoutas & Nikiforakis 2012; Balafoutas et al. 2014; Yamagishi 1988). Therefore it is perhaps unsurprising that in real-life we rarely observe commuters assaulting fare-dodgers or tax-payers accosting defrauders (Traulsen et al. 2012). Instead, punishment is typically handed over to a centralized authority (Hobbes 1651; Baldassarri & Grossman 2011; Greif 1993; Rustagi et al. 2010; Kümmerli 2011; Boone 1992; Marlowe & Berbesque 2008; Wellington 1976). For example, 11th century merchants in Europe created guilds to enforce trade contracts (Greif 1993), hunters in rural African villages turn to their chiefs to settle disputes (Gibson & Marks 1995) and members of trade unions turn to their union to discipline strike-breakers (Wellington 1976). In modern times these central authorities often take the form of institutions, such as the police, who we pay (e.g. through taxes) to punish those who violate the laws of the society. Yet centralized punishment is costly even if there are no cheats to be punished - in the same way as the maintenance of a police force causes costs even if no crimes are committed - and in this sense centralized punishment is less efficient than decentralized punishment (Sigmund et al. 2010; Sigmund et al. 2011). However, theoretical and laboratory studies have shown that centralized punishment can increase cooperation and average payoffs (Kube & Traxler 2010; Boyd et al. 2010; Steiner 2007; Sigmund et al. 2010; Baldassarri & Grossman 2011; Falkinger et al. 2000) and that when those who don't contribute towards paying for centralized punishment are themselves punished (e.g. the punishment of tax dodgers), contributors in centralized punishment systems do better than punishers in decentralized systems (Sigmund et al. 2010; Sigmund et al. 2011; Zhang et al. 2013). Moreover, in experimental games subjects have been shown to prefer to pay a centralized authority to mete out punishment on their behalf, rather than to punish cheats themselves (Traulsen et al. 2012; Andreoni & Gee 2011).

Other anthropological studies have suggested that even when punishment is not centralized, punishment might only occur if a critical mass of peers agree to participate in punishing the cheat (Mathew & Boyd 2011; Wiessner 2005; Mahdi 1986; Boehm 1993). For example, warriors of the pastoralist Turkana society of

Kenya often administer corporate punishment and fines on individuals that desert during raids on neighbouring ethnic groups. Importantly, punishment only occurs if a sufficient number of peers assemble to punish the cheat (Mathew & Boyd 2011). Consistent with these anthropological studies, experimental and theoretical work has shown that coordinated punishment can raise group average payoffs (Ertan et al. 2009; Boyd et al. 2010). Nevertheless, while some real-world evidence of human punishment does exist, none of these studies have explicitly examined whether punishment induces cooperation from the target in future interactions.

1.3 When does punishment promote cooperation?

Punishment has been studied most extensively in the PGG by introducing a second stage in which players are able to pay a fee to impose a fine on their peers (e.g. Ostrom et al. 1992; Fehr & Gächter 2000; Fehr & Gächter 2002; Egas & Riedl 2008; Sefton et al. 2007; Bochet et al. 2006; Page et al. 2005; Yamagishi 1986). In the context of a PGG, cheating (often referred to as 'free-riding', (Fehr & Gächter 2000; Fehr & Gächter 2002; Egas & Riedl 2008; Yamagishi 1986) is typically defined as any contribution that is less than the population mean. Despite the cost incurred by punishers, laboratory studies have found that humans are often willing to invest to punish free-riders (Ostrom et al. 1992; Fehr & Gächter 2000; Fehr & Gächter 2002; Egas & Riedl 2008; Sefton et al. 2007; Bochet et al. 2006; Page et al. 2005; Yamagishi 1986; Dreber et al. 2008; Nikiforakis & Normann 2008; Denant-Boemont et al. 2007; Gürerk et al. 2006; Rockenbach & Milinski 2006). Moreover, in a laboratory experiment, individuals have been shown to choose to join groups where punishment occurs over groups where punishment was not possible (Gürerk et al. 2006).

A slew of studies have shown that in a PGG setting punishment does increase cooperation in comparison to when punishment is not possible (Yamagishi 1986; Ostrom et al. 1992; Fehr & Gächter 2000; Fehr & Gächter 2002; Egas & Riedl 2008; Sefton et al. 2007; Bochet et al. 2006; Page et al. 2005). For instance, the seminal studies by Fehr & Gächter (2000; 2002) demonstrated that, in the absence of punishment, contributions to a public good started reasonably high (players

contributed on average around 60 % of their endowments) but decreased drastically after multiple rounds. However, they found that when punishment was possible, contributions started higher than in the no-punishment condition and increased rapidly to almost 100% (Fehr & Gächter, 2000; 2002).

Previous studies have suggested that the use and effectiveness of punishment is dependent on the efficiency of the punishment technology available (Nikiforakis & Normann 2008; Egas & Riedl 2008; Falk et al. 2005; Vukov et al. 2013). The efficiency of punishment can be described by the fee to fine ratio, whereby a ratio of 1: 3 would reflect punishment that cost the punisher 1 unit for every 3 units removed from the targets payoff. Several studies have shown that the more efficient punishment is (i.e. the cheaper punishment is to administer), the more it is used (Nikiforakis & Normann 2008; Ostrom et al. 1992; Falk et al. 2005; Egas & Riedl 2008; Nikiforakis et al. 2010) and the more effective punishment is at sustaining cooperation (Nikiforakis & Normann 2008; Egas & Riedl 2008; Falk et al. 2005; Vukov et al. 2013). An empirical study that investigated the use and effectiveness of a range of fee to fine ratios found that punishment with a fee to fine ratio of 1: 2 or lower is required to prevent cooperation from deteriorating over time (Nikiforakis & Normann 2008). However, despite the findings of these individual studies, a meta-analysis revealed that there was no difference in the effect of punishment with various fee to fine ratios (1 : 1, 1 : 2, 1: 3 and 1 : 4) on cooperation (Balliet et al. 2011), suggesting that the effect the efficiency of punishment on its use and effectiveness remains unclear.

Although higher cooperation levels typically imply higher payoffs for the group, this is not necessarily true if players are allowed to invest in punishment because punishment is costly to both the punisher and the target. While punishment (the availability of) ostensibly increases cooperation levels in the PGG setting (Yamagishi 1986; Ostrom et al. 1992; Fehr & Gächter 2000; Fehr & Gächter 2002; Egas & Riedl 2008; Sefton et al. 2007; Bochet et al. 2006; Page et al. 2005), several studies have shown that punishment has no effect (Bochet et al. 2006; Page et al. 2005) or even reduces group payoffs (Egas & Riedl 2008; Fehr & Gächter 2002; Ostrom et al. 1992; Sefton et al. 2007) in all but long-run

encounters (Gächter et al. 2008). In addition, although relatively few studies explore the effect of punishment in the PD game, one such study found that while punishment did increase cooperation levels in a PD game, punishment reduced payoffs at the group level and perhaps most interestingly punishers received lower payoffs than non-punishers (Dreber et al. 2008).

Punishment is especially likely to reduce payoffs to punishers where it provokes retaliation (e.g. Denant-Boemont et al. 2007; Fehl et al. 2012; Janssen & Bushman 2008; Engelmann & Nikiforakis 2012). Retaliation refers to the willingness of an individual to avenge punishment with counter-punishment. Retaliation increases the costs associated with punishing because, in addition to the cost of inflicting the punishment, the punisher also incurs the cost of being reciprocally punished. An empirical study that investigated the effect of retaliation on the effectiveness of punishment found that one quarter of all punishments were retaliated and that the threat of retaliation made individuals less likely to invest in punishment leading to a breakdown of cooperation (Nikiforakis & Normann 2008). When Engelmann & Nikiforakis (2012) allowed retaliation they found that, punishment reduced payoffs even in long-run encounters. These findings have led colleagues to question whether punishment is effective at sustaining cooperation (Dreber et al. 2008), or perhaps evolved in some other context such as enforcing dominance hierarchies.

The majority of previous work investigating the effect of punishment on cooperation has made the unrealistic assumption that all individuals are equal. In reality, however, individuals are likely to vary in their strength or resource holding potential. This variation is likely to translate into asymmetries in players' ability to punish one another and the formation of dominance hierarchies. Punishment is expected to operate down these dominance hierarchies, with dominants more likely to punish subordinates who are, in turn, unlikely to retaliate (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012). This prediction is supported by experiments using cleaner fish (*Labroides dimidiatus*), which have shown that males (the larger sex) punish females that cheat during joint inspections of model clients but—apparently due to the size difference; females never retaliate or

punish males (Raihani et al. 2010; Raihani, Grutter, et al. 2012). Power asymmetries may also fundamentally affect the outcome of human interactions involving punishment. Nevertheless, power asymmetries have not been investigated in (i) 2 player games or (ii) when retaliation is possible.

1.4 The proximate motivations underpinning punishment

Empirical studies have demonstrated that as well as the victims of cheats ('second-parties'), third-parties are also often prepared to invest in punishing cheats, even though they are not directly affected by a cheat's behaviour (known as 'third-party punishment'; Fehr & Gächter 2002; Gintis 2000; Fehr et al. 2002; Gintis et al. 2003; Fehr & Fischbacher 2003; Fowler 2005; Mathew & Boyd 2011; Charness et al. 2008; Carpenter et al. 2004). Nevertheless, the proximate motivations underpinning second-party and third-party punishment are not well understood. Negative emotions – particularly anger – are thought to promote punishment of cheats by both second-parties (Fehr & Gächter 2002; Sanfey et al. 2003; Xiao & Houser 2005; Grimm & Mengel 2011; Wang et al. 2011) and third-parties (Fehr & Fischbacher 2004b; Jordan et al. 2014). Moreover, neuroimaging studies have indicated that undertaking punishment activates reward centres in the brain of second and third party punishers, suggesting that the act of administering punishment can be subjectively rewarding for both (de Quervain et al. 2004; Strobel et al. 2011; Sanfey et al. 2003; Buckholtz et al. 2008).

Nevertheless, the negative emotions associated with interacting with a cheating partner could arise from variety of possible sources (Raihani & McAuliffe 2012a). Firstly, since victims of cheats incur a reduction in payoff, negative emotions may therefore precipitate from a desire to reciprocally harm cheating partners (also termed the desire for 'revenge'; McCullough et al. 2013). Second, extensive experimental evidence suggests that disadvantageous inequality aversion (the disutility associated with experiencing lower payoffs than others in a social interaction, (Fehr & Schmidt 1999) is an important motivator for both second-party punishment (Johnson et al. 2009; Raihani & McAuliffe 2012b) and third-party punishment (Pedersen et al. 2013). This raises the possibility that punishment may be motivated by a desire for equality. However, from these

studies we cannot rule out the possibility that punishment is simply related to the disutility associated with experiencing inequality and is not tailored to achieve equal outcomes.

Alternative studies have suggested that third-party punishment is motivated by the violation of a broadly recognized group norm (i.e. descriptive norms), rather than simply by a personal aversion to cheats or disadvantageous inequality. For instance, in a PD game defectors were more severely punished by a third-party if their partner cooperated than if their partner also defected (Fehr & Fischbacher 2004b). Similarly, individuals in a PGG were more likely to be punished by a third-party the more their contribution deviated from the group average (Carpenter & Matthews 2012). However, since the costs of punishment typically increase with the number of individuals that are punished (e.g. Fehr & Fischbacher 2004; Carpenter & Matthews 2012), we cannot rule out the possibility that individuals are more likely to punish atypical defectors simply because this is by definition cheaper than punishing defectors when defection is common.

1.5 Thesis structure

I begin in Chapter 2 with a description of the methodological and statistical techniques which were used throughout this thesis. The questions this thesis addresses are as follows:

- In Chapter 3 I ask: What effect do power asymmetries have on cooperation and punishment in a repeated Prisoners Dilemma game with binary investments?

- In Chapter 4 I ask: What effect do power asymmetries have on cooperation and punishment in a repeated Prisoners Dilemma game where players are able to vary their investment in cooperation?

- In Chapter 5 I ask: Is punishment motivated by a desire for revenge or a desire for equality?

- In Chapter 6 I ask: To what extent do the motivations for punishment vary according to framing or other contextual effects?

- In Chapter 7 I ask: Can the lower cost of punishing atypical behaviour explain the punishment of descriptive norm violators?

Finally, in chapter 8 I synthesise the findings of this thesis and discuss its contribution to and implications for research on cooperation and punishment. I also suggest several avenues for future research.

1.6 Glossary

Antisocial punishment: punishment that is aimed at cooperative or non-cheating individuals.

Centralized punishment: punishment devolved from a legitimate authority.

Cheating: behaviour in which individuals do not cooperate (or cooperate less than their fair share) but still benefit from the positive interactions with cooperating individuals.

Cooperation: the outcome of a social interaction between two or more members that results in net direct fitness benefits to each player.

Decentralized punishment: punishment carried out by other members of the social group of the cheat (often referred to as peer-punishment).

Defection: behaviour that increases immediate pay offs to the defector but which reduces the immediate pay offs of their partner (a form of cheating).

Evolutionarily stable strategy (ESS): a strategy which, if adopted by a population of players, cannot be invaded by another, initially rare strategy.

Free-riding: involves contributing less than the group mean to a public good (a form of cheating).

Nash Equilibrium: a strategy chosen by players in a game of two or more players where each player knows the equilibrium strategies of the other players and no player can benefit by changing their own strategy.

Prisoner's dilemma (PD) game: an experimental game involving two players who must both decide whether to cooperate or defect. Defection always yields a

higher pay-off than cooperating irrespective of the partner's behaviour. The prisoners' dilemma games can be a one-shot game or can be repeated over several rounds with the same partner. A repeated prisoner's dilemma is also known as an iterated prisoner's dilemma (IPD).

Public goods game (PGG): an experimental game where players are endowed with an initial sum of money and must decide how much money to contribute to a communal pot. The communal pot is then multiplied by a factor (greater than one and less than the number of players, N) and then divided among all players in the game, regardless of how much they contributed. In these games, the most profitable strategy is to withhold contributions and 'free-ride' on the investments of others. The public goods game can be a one-shot game or can be repeated over several rounds with the same group members.

Punishment: occurs when an individual reduces their own payoff to harm a cheating partner.

Second-party punishment: refers to a scenario where a cheating individual is punished by an individual who was affected by the cheat's behaviour.

Third-party punishment: refers to a scenario where a cheating individual is punished by an uninvolved bystander.

Direct fitness: lifetime reproductive success (i.e. the number of total offspring that survive until adulthood).

Indirect fitness: component of one's fitness that is increased by helping relatives to reproduce and in doing so, helping shared genes get passed on to the next generation.

Chapter 2

General Methods

2.1 Recruitment of subjects via Amazon Mechanical Turk

In Chapter 5, 6 & 7 of this thesis, I recruited subjects to take part in experimental games using the online labour market, Amazon Mechanical Turk (MTurk; www.mturk.com). MTurk connects 'requesters' (or experimenters) with 'workers' (or subjects), the latter being incentivized to perform short tasks for small payments. MTurk workers are identified by a unique 14-digit code rather than their names. Workers were told that their ID would not be revealed to their partner in the game, thus ensuring anonymity. Workers were prevented from participating repeatedly in the experiment by allowing only one entry per unique ID. Worker IDs must be linked to a valid credit card, which largely prevents workers from accruing multiple accounts (Horton et al. 2011). Subjects recruited on MTurk were redirected to an external survey website (https://opinio.ucl.ac.uk) to take part in the experiment. As a further step to prevent subjects from taking part in experiments more than once, the external survey website prevented repeated access from the same IP address.

Players' payoffs were determined by collecting the decisions of all subjects, and then once the experiment was over, matching them with a partner ('ex-post matching'; as in Horton et al. 2011). This means that players were told how their partner behaved (e.g. whether or not their partner stole from them) before actually being matched with another player. Using this method meant that subjects who interacted with each other did not need to be present at precisely the same time and no sophisticated software for simultaneous play was needed. When possible, players were matched with a unique partner in the same treatment/punishment condition/scenario as them self. However, because we did not know how players would behave (e.g. whether or not they would steal from there partner) in each treatment/punishment condition/scenario, this was not always possible. Thus, players were occasionally matched with a unique partner within a different treatment/punishment condition/scenario as them self. At the end of this process I was typically left with a few players whose behaviour did not match the behaviour described to their partners. For example, I was left with players that were told that their partner had stolen and partners that did not steal or *vice versa*. When this

occurred these players were matched with partners that had already been matched with another player. The payoff of players who were matched with more than one partner was based on only one of their matches.

Although several studies have demonstrated the validity of MTurk as a tool for collecting behavioural data (Horton et al. 2011; Rand 2012; Suri & Watts 2011), there are potential issues that experimenters have to be aware of when using this online platform for behavioural research. An advantage of using MTurk over typical western, educated, industrialized, rich and democratic (WEIRD; Henrich et al. 2010) samples is that MTurk allows for recruitment of a more diverse demographic sample (Buhrmester et al. 2011) and for subjects from non-Western world cultures to also be included so that cross-cultural effects on behaviour can be explored (e.g. Raihani et al. 2013). Nevertheless, experimenters relinquish a degree of control over the experimental setting since they cannot be certain that subjects complete the task alone (although most report that they do; Chandler et al. 2014) and are not distracted by performing other tasks (e.g. instant messaging) simultaneously (Chandler et al. 2014). The use of attention checks, built in as comprehension questions, can be used to screen out subjects who either do not attend to or do not understand the nature of the task (Goodman et al. 2013).

Although the hourly rates of MTurk subjects are usually comparable with those in the laboratory, concerns remain regarding whether the smaller stakes typically used in MTurk experiments systematically affect the decisions made by subjects. Previous laboratory studies have found that in both trust games (Johansson-Stenman et al. 2005) and in public goods games with punishment (Kocher et al. 2008), players' decisions were unaffected by increasing the stake size. Similarly, studies using dictator games have found no effect of stake size on dictator donations either in the laboratory (Forsythe et al. 1994; Carpenter et al. 2005; Cherry et al. 2002; Hoffman et al. 1996) or on MTurk (Raihani et al. 2013). However, the affect of stake size on behaviour in the ultimatum game are more mixed. Whilst, Forsythe et al. (1994) found that doubling the stake size (from $5 to $10) had no effect on either the size of offers made by proposers or on the rate at which responders rejected low offers, alternative studies have found that more

extreme increases in stake size can effect players' behaviour (Hoffman et al. 1996; Cameron 1999; Andersen et al. 2011; Munier & Zaharia 2002; Slonim & Roth 1998). For example, in a study conducted in North East India, Anderson et al. (2011) varied stake sizes from 20 Rupees (equivalent to 1.6 hours work) to 20,000 Rupees (equivalent to 1600 hours work) and found that as stakes went up, proposers offers to responders decreased and responders were less likely to reject proportionally equivalent offers.

Perhaps most concerning is the finding that the MTurk subject base has become increasingly experienced with common behavioural experiments over time; and that performance in some tasks has been shown to vary with the level of experience (Rand et al. 2012). Specifically, in a study conducted on MTurk, Rand et al. (2012) showed that decisions made under time pressure varied systematically with subject experience. Despite this fact, other studies have shown that there are no systematic differences in the responses of experienced versus naive subjects when playing common economic games (Raihani & Bshary, in review) and that subjects display remarkable consistency in responses, both across different games used to measure cooperative tendency conducted on MTurk and in self-reports of similar behavioural measures in real-life (Peysakhovich et al. 2014). On balance, therefore, I feel that so long as appropriate measures are taken to exclude subjects who do not pay attention to or do not understand the task, then MTurk should yield results that are comparable to those that could be obtained using other experimental settings.

2.2 Model averaging

In Chapters 3, 4, 6 & 7, to determine the importance of explanatory terms in my models, I employed an information theoretic approach with model averaging (as described by Grueber et al. 2011). Under an information-theoretic approach, a series of candidate models are generated, with each model representing a biological hypothesis. Rather than testing a null hypothesis, the relative degree of support for each model from the candidate set is calculated (Burnham & Anderson 2002). Initially, I specified a global model which included all the explanatory terms specified above. The input variables were centred by subtracting the mean

(Schielzeth 2010), centring allows averaging over models that include different interaction terms (Grueber et al., 2011). Although for binary variables this mean value does not exist in reality, when the mean is subtracted from a binary variable the difference between the levels for the two categories will still be 1 (e.g. −0.25 and 0.75), this means that estimates of binary explanatory terms still express the expected change compared with the reference category. Centering input variables leads to explanatory terms and the intercept being estimated at the other explanatory variables mean value. For binary variables, estimating the other explanatory terms at this imaginary mean level makes sense because these estimates can be interpreted as the average effect across different levels of the binary predictor (Gelman 2008; Grueber et al. 2011; Schielzeth 2010). After centering, continuous input explanatory variables were then standardized by dividing by 2 standard deviations (Gelman, 2008). Standardization allows the relative strength of parameter estimates to be interpreted (Gelman, 2008).

I used the package MuMIn (Barton 2013) to derive and compare submodels from the initial global model. Models were compared to one another using Akaike's Information Criterion corrected for small sample sizes (AICc; Hurvich & Tsai 1993). A subset of 'top models' were defined by taking the best model (the model with the lowest AICc value) and any models within 2AICc units of the best model (following Burnham & Anderson 2002). Using this subset of 'top' models, I computed the average parameter estimates for each term included in the subset of models, as well as the relative importance of the term. Importance is calculated by summing the Akaike weights of all models where the term in question is included in the model. Akaike weights represent the probability of a given model being the true model (compared to other candidate models in the set; Burnham & Anderson 2002). Importance can therefore be thought of as the probability that the term in question is a component of the best model (Symonds & Moussalli 2011). In the results sections of each chapter, I only present the parameter estimates from the top models (those that were within 2 AICc units of the best model).

Chapter 3

The Effect of Power Asymmetries on Cooperation and Punishment in a Prisoner's Dilemma Game

3.1 Note

This work has been published as Bone JE, Wallace B, Bshary R, Raihani NJ (2015) doi: 10.1371/journal.pone.0117183. Nichola Raihani contributed to experimental design and discussion. Redouan Bshary contributed to discussion. Brian Wallace contributed to data collection and discussion. I designed the experiment, collected the data, analysed the data and wrote the paper.

3.2 Abstract

Recent work has suggested that punishment is detrimental because punishment provokes retaliation, not cooperation, resulting in lower overall payoffs. These findings may stem from the unrealistic assumption that all players are equal: in reality individuals are expected to vary in the power with which they can punish defectors. Here, we allowed strong players to interact with weak players in an iterated prisoner's dilemma game with punishment. Defecting players were most likely to switch to cooperation if the partner cooperated: adding punishment yielded no additional benefit and, under some circumstances, increased the chance that the partner would both defect and retaliate against the punisher. Our findings show that, in a two-player game, cooperation begets cooperation and that punishment does not seem to yield any additional benefits. Further work should explore whether strong punishers might prevail in multi-player games.

3.3 Introduction

Punishment involves an individual paying a cost in order to inflict harm on a cheat (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012). This investment can be recouped if the punished individual - or a bystander - behaves more cooperatively in subsequent encounters (Clutton-Brock & Parker 1995; Raihani et al. 2010; dos Santos et al. 2013). Although people are apparently willing to pay to harm cheats (Bochet et al. 2006; Botelho et al. 2005; Denant-Boemont et al. 2007; Egas & Riedl 2008; Fehr & Gächter 2002; Gürerk et al. 2006; Nikiforakis & Normann 2008; Ostrom et al. 1992; Page et al. 2005; Rockenbach & Milinski 2006; Yamagishi 1986) and even derive subjective pleasure from doing so (de

Quervain et al. 2004; Buckholtz et al. 2008), consensus is still lacking over whether punishment is effective at promoting cooperation (Fehr & Rockenbach 2003; Houser et al. 2008; Fehr & Gächter 2000; Gürerk et al. 2006; Sigmund 2007; Yamagishi 1986; Fehr & Gächter 2002; Vukov et al. 2013). Although punishment increases payoffs in long-run encounters (Gächter et al. 2008), others have found no benefit to punishers (Bochet et al. 2006; Botelho et al. 2005; Page et al. 2005), or that punishment reduces the payoffs of the punisher or their group (Dreber et al. 2008; Egas & Riedl 2008; Fehr & Gächter 2002; Ostrom et al. 1992; Sefton et al. 2007). Punishment is especially likely to reduce payoffs to punishers where it provokes retaliation, rather than cooperation, because, as well as the cost of inflicting the punishment, the punisher incurs an additional cost of being punished (e.g. Dreber et al. 2008; Nikiforakis 2008; Janssen & Bushman 2008; Fehl et al. 2012). Indeed, when retaliation is possible, punishment has been shown to reduce payoffs even in long-run encounters (Engelmann & Nikiforakis 2012). Based on these findings, it has been argued that rewards are more effective at sustaining cooperation (Rand et al. 2009) and that punishment is unlikely to have evolved as a cooperation-enforcing mechanism (Dreber et al. 2008).

Some of the puzzling findings regarding punishment might stem from the assumption in several studies (e.g. Dreber et al. 2008; Egas & Riedl 2008; Fehr & Gächter 2002) that all players are equal. In reality individuals are expected to often vary in power or resource holding potential, such that some players are able to inflict a greater cost on the partner than the partner is able to reciprocate. In fact, it has been suggested that punishment is most likely to operate down a dominance hierarchy, with dominants punishing subordinates who are, in turn, unlikely to retaliate (Clutton-Brock & Parker 1995; Axelrod 1984; Raihani, Thornton, et al. 2012; Wang et al. 2010; Bshary et al. 2008). For example, experiments using cleaner fish (*Labroides dimidiatus*), have shown that males (the larger sex) punish females that cheat during joint inspections of model clients but – apparently due to the size difference - females never retaliate or punish males (Raihani et al. 2010; Raihani, Grutter, et al. 2012). Similarly, a recent field study revealed that both human males and females were more likely to confront females than males for dropping litter. Moreover, fear of retaliation was the most common

reason survey respondents gave for not confronting litterers (Balafoutas et al. 2014). Power asymmetries might therefore be expected to fundamentally affect the outcome of human interactions involving punishment but have received relatively little attention in comparison to other asymmetries (e.g. in cost of contribution, Tan 2008; Noussair & Tan 2011), or benefits derived from (Reuben & Riedl 2013; Janssen et al. 2011; Nikiforakis et al. 2012) a public good). One study which did explore the effect of power asymmetries showed that in the setting of a public goods game, strong players contributed similar amounts as weak players but also punished more and received higher payoffs than weak players (Nikiforakis et al. 2010). Although punishment raised the contributions of low contributors, this study did not explore whether the effectiveness of punishment was affected by power asymmetries, as would be expected. Furthermore, players were not informed which of their peers punished them; thereby preventing retaliation. Thus, the effect of power asymmetries on punishment use and effectiveness in humans remains poorly understood.

Here, we incorporated power asymmetries into a two-player iterated prisoner's dilemma (IPD) game (Luce & Raïffa 1957) with punishment (similar to Dreber et al. 2008) in order to explore how asymmetries affected the use and effectiveness of punishment in a setting where retaliation was possible. We used a two-player (rather than multi-player) game in this experiment as this allowed us to study the interaction between weak and strong players without the confounding effect of multiple other players' behaviour. Asymmetries were incorporated into the game by allowing strong players to interact with weak players. As in, Nikiforakis et al. (2010), investing in punishment cost all players the same amount but strong players could inflict greater damage through punishing than weak players. Several studies have shown that punishment which inflicts greater damage is used more frequently (Nikiforakis & Normann 2008; Ostrom et al. 1992; Falk et al. 2005; Egas & Riedl 2008; Nikiforakis et al. 2010) and is more effective at promoting cooperation (Nikiforakis & Normann 2008; Egas & Riedl 2008; Falk et al. 2005; Vukov et al. 2013) than milder punishment. We predicted that weaker players would be more cooperative in asymmetric than symmetric games, particularly after being punished for defecting. We expected weak players would rarely punish

in asymmetric games and that power asymmetries would reduce the likelihood that weaker players retaliate in response to punishment, thereby decreasing the cost associated with punishment for stronger players (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012; but see Nikiforakis et al. 2012).It was envisaged that by these mechanisms, punishment use for strong players could promote cooperation more effectively in asymmetric games than in symmetric games.

3.4 Methods

3.4.1 Experimental protocol

This research was approved by the University College London ethics board (project number 3720/001). All subjects remained anonymous so informed consent about the use of personal data was deemed unnecessary and was therefore waived by the University College London ethics board. The experiment took place over six sessions, one in May 2012, one in Nov 2012, two in March 2013 and two in April 2014 in the experimental laboratory in the Department of Economics, University College London. The lab consists of twenty computers, which are visually partitioned. A total of 120 participants (63 women, 57 men, mean age ± se = 22.0 ± 0.39 years) were recruited to play an IPD game with a punishment option Players interacted anonymously in pair-wise encounters by means of computer screens using the z-Tree (Fischbacher 2007) software. All players were paid a £5.00 show-up fee. Each player played two games and their final score was summed over both games and multiplied by £0.06 to determine additional earned income. Thus, one game unit corresponded to £0.06. To allow for negative incomes while maintaining the £5.00 show-up fee, all players began each game with 75 units (£4.50) to play with. The average payment per player was £16.69 and the average session length was 90 minutes. Prior to the experiment, each player was given written instructions about the game structure and required to answer nine comprehension questions to verify their understanding of the game (see supplementary material for experimental instructions and questions). The average score from the comprehension questions was 95 %. Players were informed of the correct answers after the test.

The IPD game lasted 50 rounds. To avoid end effects (Rapoport & Dale 1966), players were told that each game would last between 20 and 100 rounds. Players' behaviour did not change abruptly towards the end of the game (Figure 3.S1 in Supporting information S2), indicating that end effects were absent. As in Dreber et al. (2008), we constructed our IPD such that cooperation implied paying a cost for the other person to receive a benefit, whereas defection implied taking something away from the other person (see Table 3.1).

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(1, 1)	(-2, 3)
	Defect	(3, -2)	(0, 0)

Table 3.1 Payoffs accruing to (Player 1, Player 2) in step 1.

Players were randomly split into two types: weak and strong. Weak players punished with a 1:1 fee to fine ratio, meaning that if they chose to punish their partner it would cost them one unit and it would also cost their partner one unit. Strong players punished with a 1:4 fee to fine ratio (as in Dreber et al. 2008), meaning that punishing their partner would cost them one unit but it would cost their partner four units. Each player played one game with a partner of the same type as themselves (symmetric) and one game with a partner of a different type (asymmetric). The order in which players played symmetric and asymmetric games was counter-balanced.

In this study we were interested in the effects of power asymmetries on punishment, rather than on coercive behaviour. Coercion is similar to punishment since they are both aggressive behaviours that can induce cooperation from the target. However, coercion differs from punishment because unlike punishment which is typically aimed at producing mutual cooperation, coercion forces the target into a position where they would do better if they could avoid interacting with the aggressor altogether, but are constrained to do so (Raihani, Thornton, et al. 2012).

To ensure that weak partners were not being coerced into cooperating with a defecting strong partner (e.g. Nikiforakis et al. 2014), players could choose to opt out of the current round (rather than cooperating or defecting). If either player chose to opt out, step two of the current round was skipped and the next round then began as normal. After both players made their choice, they were shown their own and their partner's choice and each player's payoffs from this step.

Each round of the game was split into two steps. In step one, both players simultaneously chose between the options of "cooperate", "defect", or "do not participate in this round" (opt out). In step two, players were given the option of whether or not to punish the partner. Thus, unlike Dreber et al. (2008), choosing to punish in our game did not imply that players must also forego the option to cooperate in that round. At the end of step 2, players were shown their own and their partner's choice and payoff from step 2, as well as the cumulative payoffs for both players for that round and their own total payoff (summed over all rounds). To prevent negative incomes, if a player's total payoff became zero or below, they were bankrupt and both they and their partner were unable to take part in the remaining rounds of the game. Only three players went bankrupt during the game (two were weak players in asymmetric games and one was a strong player in a symmetric game). At the end of the first game, players were presented with the final scores and then randomly re-matched for the second game.

In order to avoid framing effects, neutral language was used. Player types "weak" and "strong" were replaced with "type 1" and "type 2"; "cooperate" and "defect" were replaced with "option A" and "option B"; and "punish" and "don't punish"

were replaced by “option C” and “option D”. After both games had finished, all subjects were required to fill in a questionnaire to provide demographic information (see supplementary material for questions and demographic data).

3.4.2 Analyses

We asked how asymmetries affected (i) players' tendency to cooperate and (ii) players' tendency to punish their partner’s defection. Previous work has shown that cooperators are more likely than defectors to use punishment (Falk et al. 2005). Thus, in analysis (ii), we controlled for whether or not the player cooperated (see Table 3.2 for a full list of all models with response and explanatory terms). We then asked (iii) how punishing a cheating partner affected the punisher's payoff. To answer this question, we analysed how a players’ total payoffs at the end of each game was affected by the proportion of times which they punished in response to their partner defecting. This analysis was restricted to players whose partner defected at least once.

Next, we asked how players responded to being punished. We asked whether punishment affected the likelihood that a player would subsequently (iv) cooperate (v) retaliate or (vi) opt out in the next round in both symmetric and asymmetric games. We compared the likelihood that players cooperated, retaliated or opted out in round $n + 1$ after having been punished (or not) in round n. We classed a player as retaliating if they punished a cooperative partner in round $n + 1$, having been punished by that partner in round n.

Finally, we asked (vii) how defecting players responded to their partner opting out (rather than punishing) in the previous round. We compared the likelihood that players cooperated in round n+2 when their partner opted out of round n+1 with the likelihood that players cooperated in round n+1 when there partner did not opt out of round n+1.

For analyses (iv) - (vii), data were restricted to instances where players had defected in round n (i.e. the effect of antisocial punishment on players' behaviour was not measured). Players' behaviour in round $n + 1$ could also be affected by whether their partner cooperated or defected in round n. Therefore, in analyses

(iv) & (vi), we controlled for whether or not the partner cooperated in round *n*. However, since opportunities to retaliate were relatively rare (N = 107 opportunities for retaliation), in analysis (v), we did not have sufficient data to control for the effects of whether or not the partner cooperated in round *n* (see Table 3.2 for a full list of all models with response and explanatory terms).

For analyses (iii) and (iv), initially a three-way interaction was a component of the best model (Table 3.S2 in Supporting information S2) so separate models were produced for weak and strong players for ease of interpretation (Table 3.2; Table 3.6). A three-way interaction was also a component of the best model when data were restricted to strong players so separate models were produced for strong players with defecting partners and strong players with cooperating partners (Table 3.2). For analysis (v), a model was only produced for players with strong partners because we did not record any instances of retaliation against weak partners (Table 3.2; Table 3.3).

Model	Question	Response term	Explanatory terms	n for analysis
(i)	Do asymmetries affect cooperation?	Player defected (0) Player cooperated (1)	Player type (strong) Game type (asymmetric) Player type x game type	9610 rounds
(ii)	Do asymmetries affect punishment?	Player did not punish defecting partner (0) Player did punish defecting partner (1)	Player type (strong) Game type (asymmetric) Player cooperated (yes) All 2-way interactions and the 3-way interaction	2173 rounds
(iii)	Does punishing a cheating partner affect the punishers total payoff ?	Players total score at the end of the game	Proportion player punished Game type (asymmetric) Proportion player punished x Game type	
(a)	Player is weak			92 players
(b)	Player is strong			83 players
(iv)	Do asymmetries affect whether punishment promotes cooperation?	Player continued to defect (0) Player switched to cooperate (1)	Game type (asymmetric) Partner Punished (yes) Partner cooperated (yes) All 2-way interactions and the 3-way interaction	
(a.1)	Player is weak		(Partner cooperated and the 3-way interaction are only	936 rounds
(b.1)	Player is strong & partner defected		included in model 1.1)	476 rounds
(b.2)	Player is strong & partner cooperated			236 rounds
(v)	Do asymmetries affect whether punishment provokes retaliation?	Player did not punish cooperative partner (0) Player did punish cooperative partner (1)	Player type (strong) Partner punished (yes) Player type x Partner punished	
(a)	Partner is weak			N/A
(b)	Partner is strong			176 rounds
(vi)	Do asymmetries affect whether punishment provokes opting out?	Player did not opt out (0) Player opted out (1)	Player type (strong) Game type (asymmetric) Partner punished (yes) Partner cooperated (yes) All 2-way interactions and 3-way interactions	2157 rounds
(vii)	Do asymmetries affect whether punishment promotes opting out?	Player continued to defect (0) Player switched to cooperate (1)	Player type (strong) Game type (asymmetric) Partner opted out (yes) All 2-way interactions and 3-way interactions	1394 rounds

Table 3.2 Explanatory terms: player type is a 2-level factor with levels 'weak' and 'strong'; game type is a 2-level factor with levels 'symmetric' and 'asymmetric'; player cooperated is a 2-level factor describing whether the player cooperated or defected in the current round; partner punished is a 2-level factor describing whether or not the player was punished by their partner in the previous round; partner cooperated is a 2-level factor describing whether the partner cooperated or not in the previous round; proportion player punished is a continuous variable measuring the proportion of times that the player punished if their partner defected; partner opted out is a 2-level factor describing whether or not the player's partner opted out in the previous round.

3.4.3 Statistical methods

Data were analysed using R version 2.15.2 (R Development Core Team 2011). Generalised linear mixed models (GLMMs) were used for all analyses. For all but analysis (iii), GLMMs were fit with a binomial error structure and logit link function. For analysis (iii), GLMMs were fit with a Poisson error structure and log link function. To determine the importance of explanatory terms in our models, we used an information theoretic approach with model averaging (Grueber et al. 2011; see General methods for details). GLMMs allow repeated measures to be fitted as random terms, thus controlling for their effects on the distribution of the data. Player identity was included as a random term in all models produced. For all analyses, we excluded rounds where either player was bankrupt. Explanatory input variables were centred by subtracting their mean (Schielzeth 2010). After centring, continuous explanatory input variables were then standardized by dividing by 2 standard deviations. We only present the effect sizes from the top models (see supplementary material for effect sizes for analyses (v) and (vi)).

3.5 Results

Cooperation was relatively common in all game types and when measured over the whole game strong players were typically more cooperative than weak players (Table 3.3). Cooperation levels (over the whole game) varied with game type: although weak players were more cooperative in asymmetric games than in symmetric games, strong players were less cooperative in asymmetric games than in symmetric games (Table 3.3; Table 3.4; Figure 3.1). Thus, both weak and strong players were most likely to cooperate if their partner was strong.

Although when taking the whole game into account, cooperation levels differed according to player type and game type, for all but strong players in symmetric games these differences became vanishingly small as cooperation levels converged to higher levels over the course of the game (Table 3.3; Figure 3.2a). For strong players in symmetric games, cooperation levels also increased over the course of the game but at a slower rate than for other players. This meant that while symmetric games between strong players had the highest cooperation levels

at the beginning of the game, the cooperation levels in other game types was considerably higher in the final rounds of the game (Table 3.3; Figure 3.2a).

In general, players were most likely to punish a defecting partner if they themselves had cooperated; the magnitude of this effect was larger for weak players than strong players (Table 3.3; Table 3.5; Figure 3.3). Strong players were typically more likely to punish than weak players (Table 3.3; Table 3.5; Figure 3.3). Punishment was used most often – by players of both types – in asymmetric rather than symmetric games (Table 3.3; Table 3.5; Figure 3.3). Nevertheless, punishment use generally decreased over the course of the game (Table 3.3; Figure 3.2b).

For weak players, investing in punishment appeared to have a negative effect on payoffs in both symmetric and asymmetric games (Table 3.6; Figure 3.4a). In contrast, for strong players punishing defecting partners only appeared to reduce punishers' payoffs in symmetric games; in asymmetric games strong players' payoffs were unaffected by their investment in punishment (Table 3.6; Figure 3.4b). However, for both weak and strong players the confidence intervals for this term crossed zero, meaning that there is little evidence for these effects (Table 3.6). Nevertheless, in general weak players received higher total payoffs in symmetric games than in asymmetric games (Table 3.3). Strong players on the other hand received similar total payoffs in symmetric and asymmetric games (Table 3.3). Although from our model (Table 3.6) it appears as if strong players generally received higher total payoffs in asymmetric games than in symmetric games, this is an artifact resulting from the model excluding players whose partner never defected; specifically because these players were more prevalent in symmetric games (n = 25) than in asymmetric games (n = 12) and were typically the highest scoring players.

	Weak players		Strong Players	
	Symmetric	Asymmetric	Symmetric	Asymmetric
Cooperated (over whole game)	0.70 ± 0.01	0.77 ± 0.01	0.84 ± 0.01	0.78 ± 0.01
Cooperated (last 10 rounds)	0.91 ± 0.01	0.88 ± 0.02	0.83 ± 0.02	0.87 ±0.02
Punished partner for D (Player cooperated)	0.25 ± 0.04	0.31 ± 0.04	0.39 ± 0.04	0.59 ± 0.04
Punished partner for D (Player defected)	0.07 ± 0.01	0.11 ± 0.02	0.23 ± 0.03	0.27 ± 0.02
Antisocial punishment	0.03 ± 0.00	0.04 ± 0.00	0.05 ± 0.00	0.09 ± 0.01
Punished partner (last 10 rounds)	0.02 ± 0.01	0.04 ± 0.01	0.01 ± 0.01	0.02 ± 0.01
Opted out	0.12 ± 0.06	0.19 ± 0.01	0.08 ± 0.00	0.04 ± 0.00
Retaliation	0 ± 0	0.21 ± 0.06	0.24 ± 0.07	0 ± 0
Total payoff	100 ± 2.96	91.5 ± 4.36	99.7 ± 4.32	98.5 ± 3.75

Table 3.3 The proportion of players who cooperated (± SE) over the whole game and in the last 10 rounds of the game, the proportion of rounds in which players cooperated / defected / opted out, the proportion of instances in which the partner defected (D) / cooperated ('antisocial punishment') that cooperating players & defecting players responded with punishment, the proportion of rounds in which players punished their partner in the last 10 rounds of the game regardless of whether the player or their partner cooperated or defected, the proportion of instances in which players punished their partner in round $n + 1$ (when the partner cooperated) following being punished for defecting in round n ('retaliation') and mean total payoffs. All proportions (except for 'opted out') exclude rounds in which either player opted out.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	1.41	0.21	(0.99, 1.81)	
Player type (strong)	1.18	0.41	(0.36, 2.00)	1.00
Game type (asymmetric)	-0.1	0.06	(-0.21, 0.03)	1.00
Player type x Game type	-0.99	0.12	(-1.23, -0.75)	1.00

Table 3.4 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models for the binary response term encoding whether or not players cooperated in each round of the game (player defected = 0, player cooperated = 1).

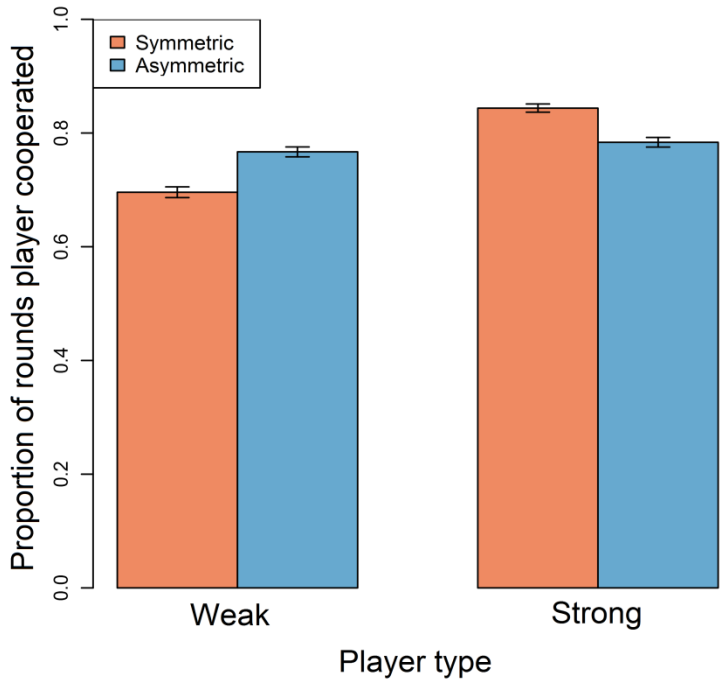


Figure 3.1 Barplot showing the proportion of rounds which players cooperated in symmetric (red bars) and asymmetric (blue bars) games according to whether they were weak or strong. Data exclude rounds where either player opted out or was bankrupt. Error bars represent standard errors. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-1.81	0.18	(-2.16, -1.47)	
Player type (strong)	1.15	0.35	(0.88, 2.22)	1.00
Game type (asymmetric)	0.88	0.19	(0.51, 1.26)	1.00
Player cooperated (yes)	1.29	0.16	(0.97, 1.61)	1.00
Game type x Player cooperated	0.47	0.29	(-0.11, 1.04)	1.00
Player type x Player cooperated	-0.75	0.31	(-1.37, -0.14)	0.65
Player type x Game type	0.16	0.36	(-0.55, 0.87)	0.19

Table 3.5 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models for the binary response term encoding whether or not players punished their partners for defecting (player did not punish defecting partner = 0, player did punish defecting partner = 1).

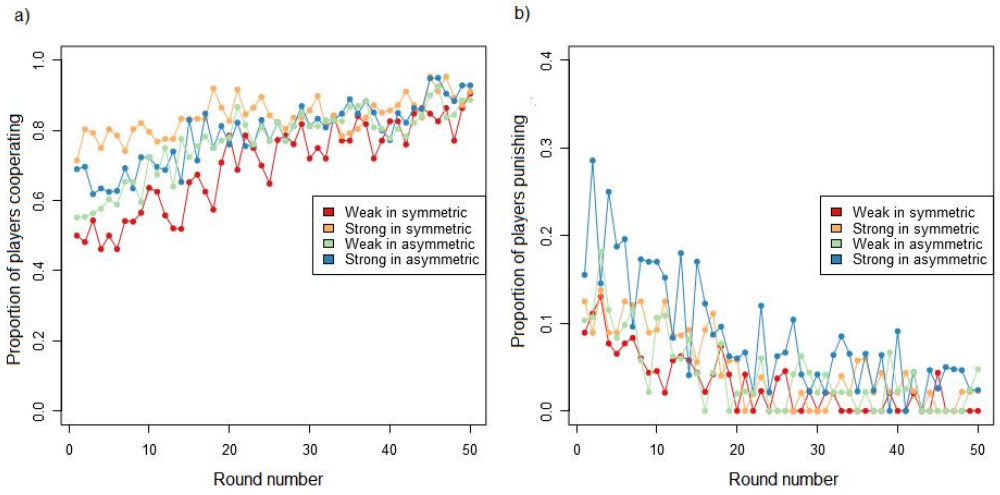


Figure 3.2 Scatter plot showing the mean proportion of players that **a)** cooperated and **b)** punished their partner, according to whether they were weak or strong and whether they were in a symmetric or asymmetric game. Rounds where either player opted out or was bankrupt were excluded.

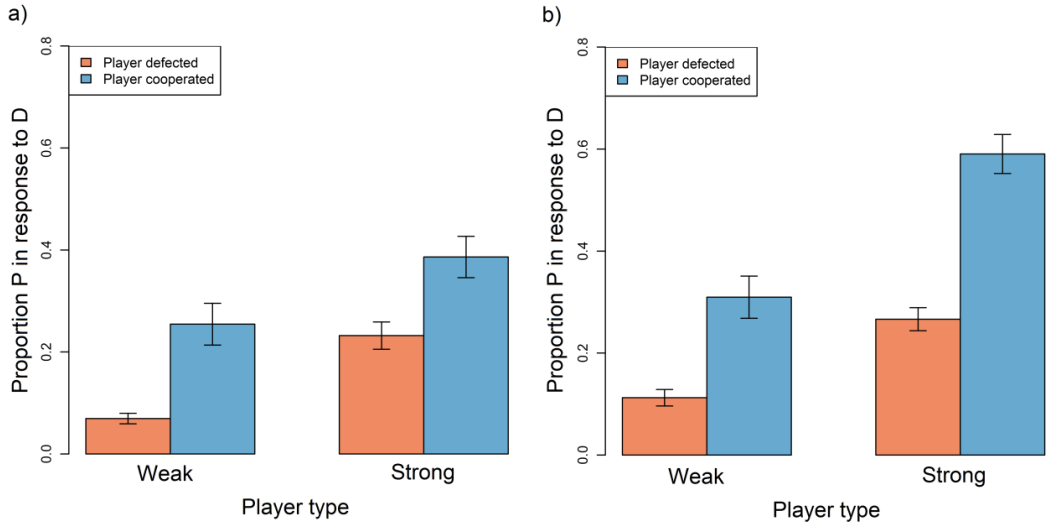


Figure 3.3 Barplot showing the proportion of their partner's defection that players punished in **a)** symmetric and **b)** asymmetric games according to whether they were weak or strong and whether they cooperated or defected them self. Data were restricted to instances in which the partner defected in the previous round, excluding rounds where either player opted out or was bankrupt. Error bars represent standard errors. Plots are generated from raw data.

Player type	Parameter	Estimate	SE	Confidence Interval	Importance
Weak	Intercept	87.1	2.98	(81.2, 93.1)	
	Game type (asymmetric)	-14.3	5.23	(-24.7, -3.86)	1.00
	Proportion player punished	-3.68	5.87	(-15.4, 7.75)	1.00
Strong	Intercept	85.0	3.69	(77.7, 92.37)	
	Game type (asymmetric)	14.4	6.53	(1.37, 27.4)	1.00
	Proportion player punished	-13.2	7.71	(-28.5, 2.12)	1.00
	Game type x Proportion player punished	27.9	14.1	(-0.18, 55.91)	0.68

Table 3.6. Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models for the response term 'total payoff'.

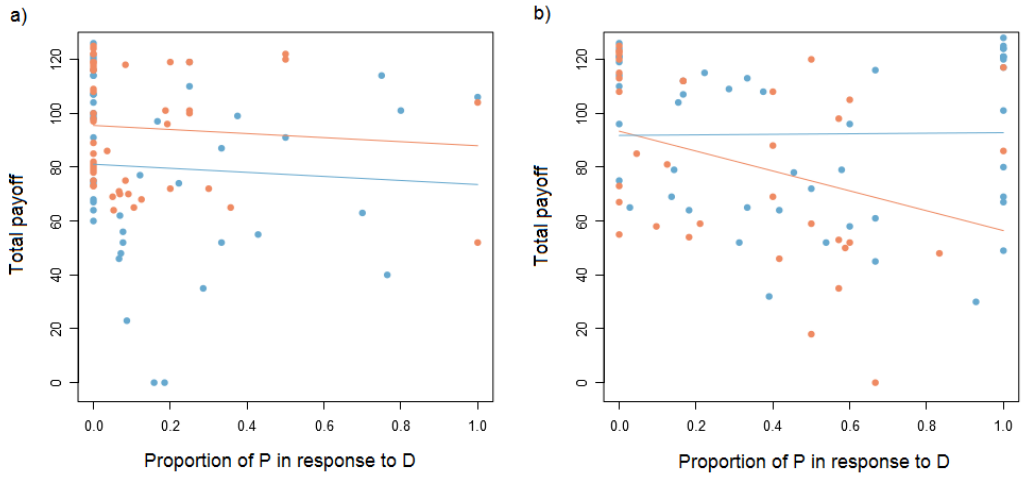


Figure 3.4. Scatter plots showing how the probability of punishing in response to defection affected total payoffs in symmetric and asymmetric games for (a) weak players and (b) strong players. Circles represent raw data points; red circles are symmetric games and blue circles are asymmetric games. Regression lines are shown for symmetric (red line) and asymmetric (blue line) games.

If the punisher had also defected (i.e. punishment was 'unjustified' or 'hypocritical'), then punishment had no meaningful effect on the target's propensity to cooperate in the next round (Table 3.7; Figure 3.5). The effect of being punished by a cooperative partner ('justified punishment') produced more variable outcomes on the target's behaviour. Where justified punishment was aimed at weak players, it produced no discernible effect on subsequent tendency to cooperate (cooperate if not punished = 0.43 ± 0.04; versus if punished = 0.35 ± 0.05). Justified punishment aimed at strong players produced different outcomes. When the punishment was administered by a weak partner, it had no discernible effect on the strong player's cooperative behaviour in the next round (cooperate if not punished = 0.59 ± 0.06; versus if punished = 0.61 ± 0.09). However, in a strong-symmetric game, justified punishment actually reduced the tendency of the player to cooperate in the next round (cooperated if not punished= 0.71 ± 0.05; versus if punished = 0.22 ± 0.06; Table 3.7; Figure 3.4). Rather than responding to punishment, defecting players were more likely to subsequently cooperate when the partner had cooperated (compared to when the partner had also defected) (cooperate if partner cooperated = 0.48 ± 0.02; versus if partner defected = 0.1 ± 0.01). Thus, cooperative behaviour from the partner appeared to be sufficient to convert defectors to cooperators, and punishment in addition to cooperation did not yield any additional benefits.

Neither weak nor strong players ever retaliated against a weak partner (Table 3.3). However, weak and strong players did and were equally likely to retaliate in response to punishment from strong partners (Table 3.3; Table 3.8; Figure 3.5). In general, weak players opted out more often than strong players (Table 3.3; Table 3.9). In addition, both weak and strong players were more likely to opt out if their partner was strong (Table 3.3) and if they were punished in the previous round (proportion of instances in which players opted out if they were not punished for defecting in the previous round ± SE = 0.11 ± 0.01 versus if they were punished for defecting in the previous round = 0.19 ± 0.02; Table 3.9). Players were also less likely to opt out if their partner cooperated, rather than defected, in the previous round (proportion of instances in which players opted out if their partner

defected in the previous round ± SE = 0.14 ± 0.01 versus if their partner cooperated in the previous round = 0.07 ± 0.01; Table 3.9)

Regardless of player type or game type, defecting players were slightly less likely to switch to cooperate if their partner opted out in the previous round. However, the confidence intervals for this term include zero, meaning that this effect is weak (proportion of instances in which players switched to cooperate if their partner did not opt out of the previous round = 0.14 ± 0.04 versus if their partner did opt out in the previous round 0.2 ± 0.01; Table 3.10).

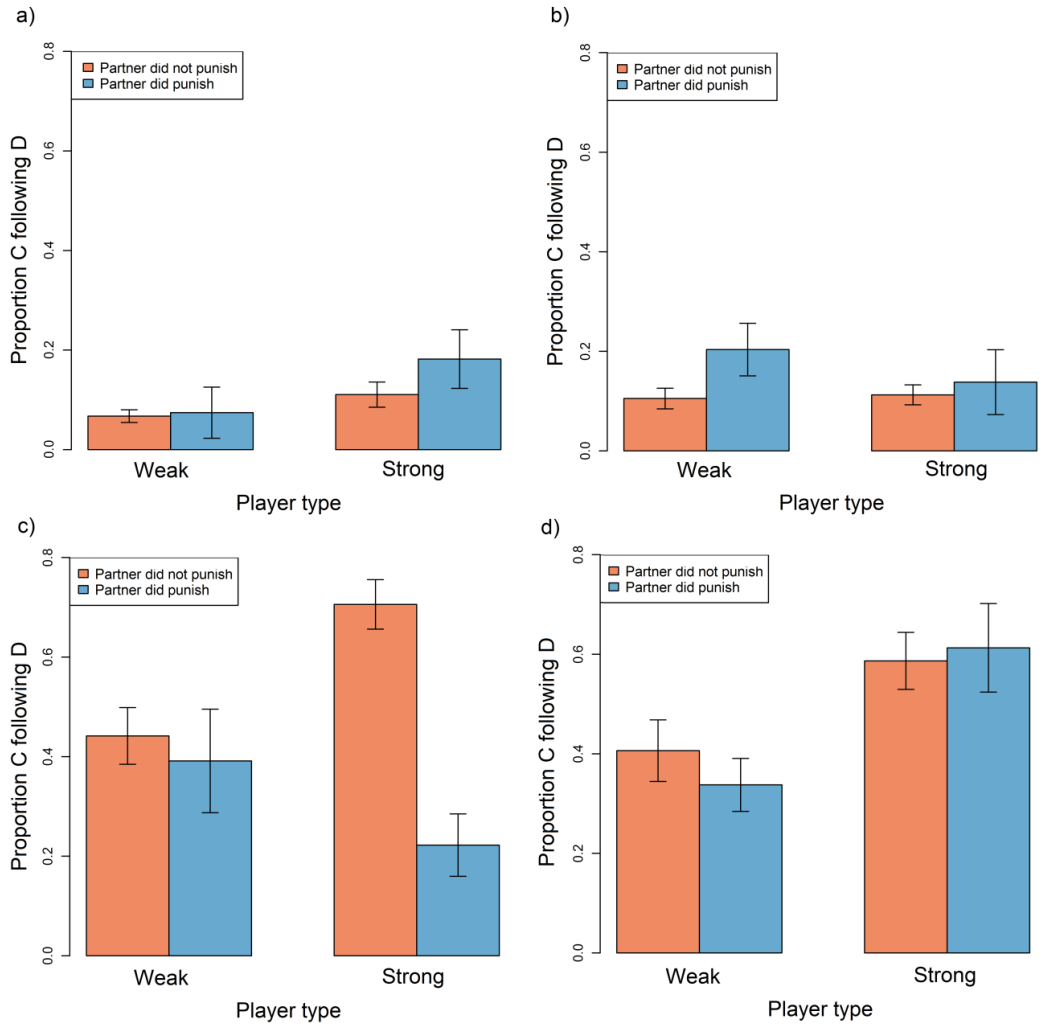


Figure 3.5 Barplot showing the proportion of instances in which players cooperated following defecting in the previous round, according to whether they were weak or strong and whether they were punished by their partner in the previous round. Data is shown for players in **a)** symmetric games where the partner defected in the previous round **b)** asymmetric games where the partner defected in the previous round **c)** symmetric games where the partner cooperated in the previous round and **d)** asymmetric games where the partner cooperated in the previous round. Data were restricted to instances in which the player defected in the previous round, excluding rounds where either player opted out or was bankrupt. Error bars represent standard errors. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-1.81	0.18	(-2.16, -1.47)	
Player type (strong)	1.15	0.35	(0.88, 2.22)	1.00
Game type (asymmetric)	0.88	0.19	(0.51, 1.26)	1.00
Player cooperated (yes)	1.29	0.16	(0.97, 1.61)	1.00
Game type x Player cooperated	0.47	0.29	(-0.11, 1.04)	1.00
Player type x Player cooperated	-0.75	0.31	(-1.37, -0.14)	0.65
Player type x Game type	0.16	0.36	(-0.55, 0.87)	0.19

Table 3.7 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models for the binary response term encoding whether or not players punished their partners for defecting (player did not punish defecting partner = 0, player did punish defecting partner = 1).

Partner type	Parameter	Effect size	SE	Confidence Interval	Importance
Strong	Intercept	-3.34	0.84	(-7.45, -2.18)	
	Punished (yes)	1.96	0.73	(0.58, 3.57)	1.00

Table 3.8 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating whether a player retaliated against a punitive partner (player did not punish cooperative partner = 0, player did punish cooperative partner = 1).

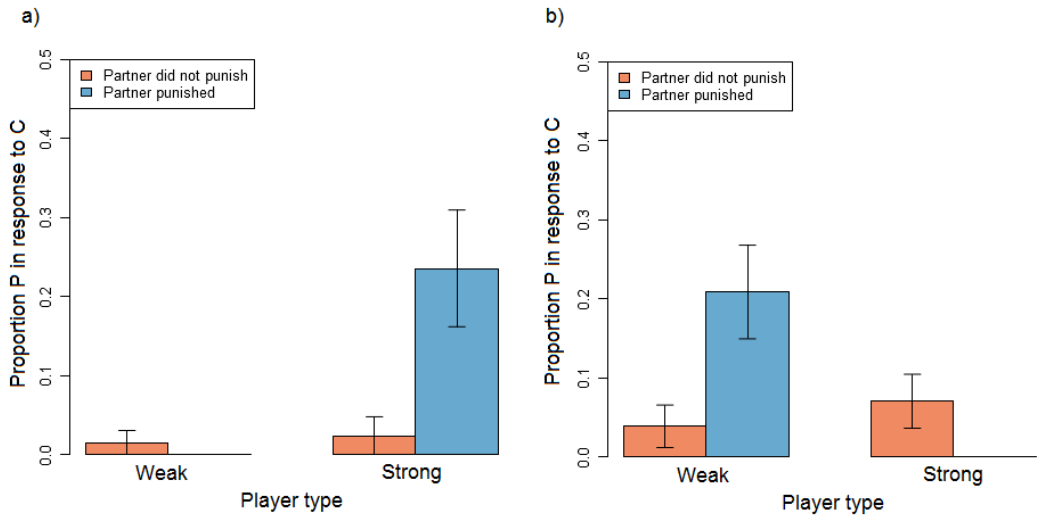


Figure 3.6 Barplot showing the mean proportion of players in **a)** symmetric and **b)** asymmetric games that punished their partner for cooperating, according to whether they were weak or strong and whether they were punished by their partner in the previous round. Data were restricted to instances in which the player defected in the previous round and their partner cooperated in the current round. Rounds where either player opted out or was bankrupt were also excluded. Error bars represent standard errors. Thus, red bars represent antisocial punishment whereas blue bars can be interpreted as retaliation for punishment previously received. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-2.60	0.18	(-2.95, -2.25)	
Player type (strong)	-1.10	0.33	(-1.74, -0.46)	1.00
Game type (asymmetric)	-0.14	0.20	(-0.54, 0.26)	1.00
Partner punished	0.94	0.20	(0.55, 1.32)	1.00
Partner cooperated	-1.04	0.24	(-1.51, -0.57)	1.00
Player type x Game type	-1.36	0.45	(-2.24, -0.47)	1.00
Player type x Partner cooperated	0.56	0.45	(-0.33, 1.44)	0.39
Game type x Partner cooperated	-0.44	0.43	(-1.28, 0.39)	0.35
Partner cooperated x Partner punished	0.38	0.44	(-0.49, 1.24)	0.31

Table 3.9 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating whether players opted out (player did not opt out = 0, player opted out = 1).

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	1.35	0.00	(-1.35, -1.35)	
Game type (asymmetric)	-0.17	0.00	(-0.17, -0.16)	1.00
Player type (strong)	1.12	0.00	(1.12, 1.13)	1.00
Game type x Player type	-1.44	0.00	(-1.45, -1.44)	1.00
Partner opted out (yes)	-0.06	0.03	(-0.06, -0.06)	0.27

Table 3.10 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating how players that defected in round n responded to their partner opting out of round n + 1 (player continued to defect = 0, player switched to cooperate = 1).

3.6 Discussion

Previous studies have suggested that punishment use is detrimental because it induces retaliation rather than cooperation (Dreber et al. 2008; Janssen & Bushman 2008; Nikiforakis 2008). However, previous studies have typically not accounted for asymmetries between players, which might affect how targets respond to being punished. In this two-player game, punishment did not promote cooperation under any circumstances. Moreover, punishment from strong players provoked (i) further defection and (ii) retaliation from the target. Therefore, punishment carried both the cost of inflicting the punishment itself and the cost associated with retaliation (for strong players) but did not confer the benefits of increased cooperation. We discuss reasons why our findings may not have matched the theoretical predictions below.

As expected, punishment was most often directed from cooperating players at defecting partners and strong players were generally more likely to punish than weak players. While these results supported our initial predictions, we had other findings that were more puzzling. First, weak players were more likely to punish in asymmetric games than in symmetric games, which is the opposite pattern to that which we predicted. Second, players seemed to respond more to the cooperative decision of the partner than to the punitive action when deciding how to behave in the next round. In situations where the partner cooperated, there was no additional positive effect of punishment on the propensity of the target to cooperate and, in fact, in some cases the punishment made the target more likely to defect. Third, while strong punishers frequently encountered retaliation from the target, we did not record a single instance of retaliation against a weak punisher. These results are clearly counter to our predictions that punishment would operate down a dominance hierarchy and be most likely to promote cooperation when directed from strong players at weak targets. In part these findings might stem from the fact that cooperating and defecting were specified as binary, all-or-nothing responses in this study, rather than continuous variables. Thus, a player that switched to cooperation (in response to the partner's cooperation) could not, by definition, increase their investment even more in

response to being punished. This is unlike the situation with cleaner fish, where, although defecting is a binary outcome (bite client / do not bite client) investment can be modelled as a continuous variable (duration of time 'cooperating' by removing ectoparasites; Bshary et al. 2008). In other real world settings it is debatable whether cooperation should be modelled as an all-or-nothing response or instead as a continuous variable (Sherratt & Roberts 1998; Wahl & Nowak 1999; Killingback et al. 1999; Killingback & Doebeli 2002). Had we specified cooperation as a continuous variable in this study, we may have been more likely to measure a meaningful effect of punishment on the target's subsequent cooperation. This remains an important avenue for further exploration.

Another possibility is that the effectiveness of punishment varies between two-player games and multi-player games. Previous theoretical and empirical studies (Dreber et al. 2008; Rand et al. 2009) have argued that in two player games selection favours strategies that cooperate conditionally rather than paying to punish cheats. Our data seem to support these arguments, even when player asymmetries are included in games. Although conditional cooperation is successful in two-player games, such strategies might be less effective in multi-player games because defecting in response to a defector harms the cooperative partners in the group as well as the cheats (Raihani, Thornton, et al. 2012). Thus, although asymmetries did not allow punishment to promote cooperation in this two-player game, they may be more effective in a multi-player game (e.g. Przepiorka & Diekmann 2013). Furthermore, punishment is likely to be most effective at promoting cooperation where the possibility for retaliation is reduced, for example when punishment is administered by a legitimate authority (Baldassarri & Grossman 2011) or when it is administered jointly by several group members (Boyd et al. 2010).

While these explanations might help us to understand why punishment did not promote cooperation, it is less clear why punishment from strong, but not weak, players was actually detrimental - being more likely to provoke both retaliation and defection. It has been suggested that the moral legitimacy of punishment is an important determinant of how the target is expected to respond (Fehr &

Rockenbach 2003; Xiao 2013). According to Fehr & Rockenbach (2003), punishment may be perceived as being morally illegitimate if it is associated with selfish or greedy (rather than altruistic) intentions. In this way, punishment that increases the payoffs of the punisher, relative to the target (as punishment from strong players in our study did), may be interpreted as a competitive act and therefore perceived as morally illegitimate. It has been argued that morally illegitimate punishment is unlikely to promote cooperation from targets (Xiao 2013; Fehr & Rockenbach 2003): an extension of this prediction might be that morally illegitimate punishment makes players more likely to defect and to retaliate. Supporting this idea, a recent study that incorporated power asymmetries in a two-player game where senior workers could exploit junior partners (by suggesting that they contribute larger investments and punishing them if they failed to obey) showed that junior workers did not obey strong partners when they knew that they were being exploited (Nikiforakis et al. 2014). Conversely, when junior workers only had incomplete information about how much the other player earned, they were more likely to comply with the senior worker's suggestion (Nikiforakis et al. 2014). Theoretical work has shown that in multi-player games, punishment may be cost-effective if multiple individuals within the group collectively punish defectors (Boyd et al. 2010). Such collective punishment may also be perceived to be more legitimate than punishment from a single group member because it is more likely to be in the collective interest, rather than in the interest of a selfish individual. Although, punishment in this study was decentralized, the legitimacy of punishment is also likely to be important under centralized punishment regimes. For example, punishment is likely to be more effective at promoting cooperation when centralized authorities have been legitimately elected; rather than chosen at random (Baldassarri & Grossman 2011). In light of these results and the results of our study, we suggest that further work to explore the moral assessment of punishment in different circumstances would be very helpful to understand why punishment sometimes promotes and sometimes undermines cooperation. It is important that this future work should collect data on both players' behavioural responses as well as their subjective evaluations of punishment.

This study suggests that punishing cheats was detrimental to punisher's payoffs (as in Dreber et al. 2008; Egas & Riedl 2008; Fehr & Gächter 2002; Ostrom et al. 1992; Sefton et al. 2007) for weak players in symmetric and asymmetric games and strong players in symmetric games, but in asymmetric games, strong players' payoffs were unaffected by punishing. However, evidence for this finding was weak. Although these results are inconclusive, if strong punishers do receive higher payoffs in asymmetric games than in symmetric games (as suggested) this may be because while strong players were equally likely to incur retaliation from weak and strong partners, retaliation from strong partners was considerably more costly than retaliation from weak partners. This effect may also be caused by the finding that while justified punishment from strong players had no effect on weak partners' tendency to switch to cooperation, it made strong partners more likely to defect in the next round. Nevertheless, in artificial laboratory settings, individual payoffs are necessarily determined by the (largely) arbitrary costs and benefits associated with the different actions available to players in the game. In addition, the time horizon of the interaction, which is again largely determined arbitrarily by the experimenter, can have a fundamental bearing on whether punishment is found to improve payoffs: punishment is least likely to be beneficial in short-run interactions (Gächter et al. 2008). Therefore, we suggest that since the net benefit of punishment in real-world settings must emerge from its ability to deter partners (or bystanders; dos Santos et al. 2013) from defecting, the effect of punishment on targets' behaviour, rather than the total payoffs accruing to the punisher, is more important for understanding the functional basis of punishment.

In general, weak players opted out more than strong players and both player types opted out more when faced with a strong partner. Although, theoretical studies have suggested that cooperation and punishment is more likely to occur and persist if players have the option to opt out (Hauert et al. 2007; Hauert et al. 2008), these models assumed that all players were equal in strength, used multi-player rather than two-player games and permitted only a small number of behavioural strategies. In our study, players were more likely to opt out of rounds if they were previously punished than when they were not punished. This suggests that both player types occasionally chose to avoid further punishment by

withdrawing from the game rather than by switching to cooperation. Punishment may be more effective at promoting cooperation when players are unable to opt out of rounds and can therefore only avoid further punishment by switching to cooperation. This hypothesis could be tested by repeating this experiment but excluding the option for players to opt out of rounds.

In this experiment, we found that cooperation levels increased over the course of the game. Our other findings suggest that this increase in cooperation was most likely to be driven by conditional cooperation rather than by punishment. Indeed, previous work has shown that cooperation in repeated prisoners dilemma games can increase over time, even in the absence of punishment (Normann & Wallace 2012). Cooperation levels of strong players in symmetric games increased at a slower rate than that of other players. We suggest that this is because punishment from strong players in symmetric games decreased cooperation rates. As cooperation levels increased with time, punishment was no longer needed so often, this lead to a decrease in punishment use over the course of the game.

Weak and strong players were also more likely to opt out of rounds if their partner defected than if their partner cooperated in the previous round. This suggests that players used opting out either to avoid defectors or to signal their disapproval with the defector's action. However, opting out appeared to decrease the probability that a defecting partner would switch to cooperation. A recent study by Delton et al. (2013) found that opting out of a public goods game was viewed as morally wrong and incited punishment from other group members. In other words, players who opted out were treated as cheats despite not actually exploiting collective benefits. The authors suggest that players who opt out may be perceived to be more likely to free-ride in the future (than those who contribute to the public good) and that punishment may therefore serve a preventative function (Delton et al. 2013). Similarly, in this study players who opted out may have been perceived to be immoral and more likely to cheat in future interactions. This may have provoked their partners to defect in order to avoid being exploited in future rounds.

To summarize, we found no evidence that punishment promoted cooperation in this two-player game and in fact punishment use sometimes had a detrimental effect on cooperation. Future work could explore whether these results stem from using a binary response term for cooperation and whether punishment might be more effective where investments are continuous. We also propose that understanding the moral assessment of punishment will be crucial for predicting when it will promote and when it will undermine cooperation.

3.7 Supplementary materials

3.7.1 Experimental instructions

Participants were given a printed copy of the instructions below. They were given 15 minutes to read through the instructions before attempting the comprehension questions. Participants were free to refer back to the instructions whilst completing the comprehension questions and throughout the experiment.

Welcome and thank you for volunteering for this experiment.

Please be quiet during the entire experiment. **Do not talk to your neighbours** and do not try to look at their screens. If you have any questions, please raise your hand. We will come to you and answer it privately.

This experiment is about decision-making. **You will be randomly assigned the role as either a "type 1" or a "type 2" player.** You will keep this role during the entire experiment. You will play a two-player game **twice.** Each game will last for a predetermined number of rounds **(between 20 and 100).**

In each game you **will be randomly matched with a different person in the room.** You will play with the same person for the duration of each game. Therefore, you will play with two people in total. In one game you will play with someone of the same type as yourself and in one game you will play someone of a different type to yourself. **You will remain anonymous throughout the experiment** and will be identified only by the name in the top left corner of your computer screen.

Before playing the games you will answer some questions about the experiment. The purpose of this is to make sure that everybody fully understands the rules of the experiment before we start. So please make sure you read these instructions carefully.

Depending on your decisions and the decisions made by the other player in each of the two games you play, you will be able to earn a considerable amount of money. The scores you receive in the game will be given in units **(1 unit= 6p).**

Everyone **will receive a show-up fee of £5.** In addition, you will be given **an additional 75 units (£4.50) to play with at the start of each game**. Depending on the decisions made by you and the other player during each game you will gain or lose units from this initial amount. After the experiment has finished, your units will be converted to real money. You will be paid the show-up fee and the money that you earn. In total this could be as much as £32.

Payment: Payment will happen after both games have been played.

The Game:

Each round will be split into two steps. In each step you will be asked to make a decision. The decisions you make will affect the score you and the other player receives for that round. There is a time limit of **15 seconds** on each step (shown in the top right corner of the screen). It is important you make a choice within this time, if not **a default choice will be picked for you** and you will move on to the next step.

Each of these steps will now be explained to you in detail:

Step 1:

Both you and the other player **simultaneously** choose between the options, **"A", "B" or "Do not participate in this round".**

If **you choose "A"** then **you** will get **−1 unit**, whereas the **other player** will get **+2 units**.

If you choose "B" then you will get +1 unit, whereas the other player will get −1 unit.

If either player chooses "Do not participate in this round" **both you and the other player will skip the second step and move to step 1 of the next round. Neither player will gain or lose any units in this entire round. The next round will begin as normal.**

Step 2: If either player chose "Do not participate in this round" in step 1 then this round is skipped. Otherwise, you and the other player are **presented with each other's choices and scores from step 1.** You must both then **decide** whether you would or would not like to reduce the other player's income at a cost to yourself by choosing between **options "C"** or **"D", respectively.**

Option "C" = reduce the other player's income at a cost of 1 unit to yourself.

Option "D" = do nothing (neither you nor the other player will gain or lose any points).

The number of units that the other player's income is reduced by when you choose option "C" varies depending upon what "type" of player you were assigned as at the start of the experiment.

Type 1 player: If you choose "C" then you will lose 1 unit, and the other player will lose 1 unit.

Type 2 player: if you choose "C" you will lose 1 unit, and the other player will lose 4 units.

After you have made a decision, you and the other player are **presented with each other's choices and incomes from step 2** as well as your **overall incomes** for the round. You will also be told your **total score** for the current game.

Your overall income in each step **is determined by the addition of the income from both your decision and the other player's decision**.

Some examples are given below;

***Step1*:**

If **both you and the other player choose "A"** then you will get **+1** (-1 from yourself, +2 from the other player = +1 total).

If **both you and the other player choose "B"** then you will get **0** (+1 from yourself, -1 from the other player= 0 total).

If **you choose "A"**, and the other player **chooses "B"** then **you will** get **–2** (-1 from yourself, -1 from the other player = -2 total).

If **you choose "B" and the other player chooses "A" then you** get **+3** (+1 from yourself, +2 from the other player = +3 total).

Step 2:** If **either player chose "Do not participate in this round" in step 1 then this step is skipped.

If you **both you and the other player choose "C"** then you will get **-2 if the other player is "type 1" player** (-1 from yourself, -1 from the other player = -2 total) **and -5 if the other player is "type 2" player** (-1 from yourself, -4 from the other player= -5 total).

If you **both you and the other player choose "D"** then you will **both get 0** (0 from yourself, 0 from the other player =0 total).

If **you choose "C"**, and the other **player chooses "D"** then **you** get -1 (-1 from yourself, 0 from the other player = -1 total).

If you **choose "D"**, and **the other player chooses "C"** then **you** get **-1 if the other player is a "type 1" player** (0 from yourself, -1 from the other player = -1 total) **or -4 if the other player is a "type 2" player** (0 from yourself -4 from the other player = -4 total).

Your income for the round will be determined by the addition of your income from step 1 and step 2.

The total number of units that you have at the end of these games will determine how much money you have earned. Therefore, the additional money

you and the other player each earn depends on which options you both choose. However, **the final scores of the other players do not matter for your earnings.**

If your **total score** drops **to 0 units or below**, you **will not be able to play** for the remaining rounds of the game of the **current game**.

At the end of both games, your total earnings will be computed. If you finish with a total score of **0 over the two games**, you will walk away with just the £5 show up fee. If you have a total score **above 0**, you will earn extra money at the exchange rate of 1 unit= 6p. The maximum extra amount that you can earn will be **£27**.

We will distribute a questionnaire at the end of the experiment that will ask some basic information about you.

Take your time to read through the instructions again. If you have any questions, please raise your hand. In a few minutes we will begin the questionnaire followed by the games.

3.7.2 Comprehension questions

All participants were required to answer the following comprehension questions before playing the game. If they answered a question incorrectly they were shown the correct answer. The possible answers are shown in parentheses.

1. Each game will last a predetermined number of rounds, between what? (10-200 / 20-100 / 0-50)

2. How many games will you play in this experiment? (1 / 5 / 2)

3. If you finish the experiment with a total score of 100 units, how much money will you earn on top of the £5 show up fee (1 unit = 6p)? (£6 / £10 / £80)

4. If your total score drops to 0 or below in a game what will happen? (You will lose your show up fee / You will have to leave the room / You will not be able to play the remaining rounds of the current game)

5. If you chose option "B" and the other player chooses option "A", how many

units will you get in step 1? (3 / 1 / 10)

6. If both you and the other player choose option "A", how many units will you get in step 1? (1 / 15 / 1)

7. If you are a "type 1" player and choose option "C" in step 2, how many units will be deducted from the other players' income? (1 / 3 / 5)

8. If both you and the other player are "type 2" players and you both choose option "C", how many units will you get in step 2? (-1 / -5 / 2)

9. If both you and the other player are "type 1" players, you both choose option "B" in step 1 and both choose option "D" in step 2, how many units will you get in that round? (0 / 2 / -5)

3.7.3 Demographic questions

After the game had finished, all participants were required to answer the following demographic questions. Responses to these demographic questions can be found in the supplementary data.

1. What is your gender?
2. How old are you?
3. What is your country of origin?
4. What is your subject of study?

3.7.4 Supplementary tables

Parameter	Weak	Strong
Age	Mean = 22.2 ± 0.65 Median = 21 IQR = 24 – 33 Range = 20 – 23	Mean = 21.8 ± 0.44 Median = 21 IQR = 23 – 33.25 Range = 20 - 23
Gender (n)	Females = 36 Males = 24	Females = 27 Males = 33
Country of Origin (n)	Canada = 1 China = 10 Cyprus = 1 France = 1 Germany = 1 Greece = 2 Hong Kong = 6 India = 4 Iraq = 2 Italy = 2 Malawi = 1 Malaysia = 3 Pakistan = 1 Poland = 3 Romania = 1 Singapore = 4 United Kingdom = 11 United States = 2 Undisclosed = 1	China = 10 Czech Republic = 1 Greece = 1 Hong Kong = 5 Hungary = 1 Malaysia = 5 Mauritius = 1 Nigeria = 1 Pakistan = 1 Romania = 2 Singapore = 8 Sri Lanka = 1 Sudan = 1 Taiwan = 1 United Kingdom = 17 United States = 1 Vietnam = 2
Subject Studied (n)	Anthropology = 1 Architecture = 2 Biochemistry = 1 Biology = 1 Biomedical Sciences = 5 Biotechnology = 1 Chemical Engineering = 1 Chemistry = 1 Cognitive Neuroscience = 1 Computer Science = 1 Economics = 10 Economics & Business = 1 Egyptian Archaeology = 1 Engineering = 3 English = 1 Genetics = 1 History of Art = 2 Humanities = 1 International Planning = 1 International Public Policy = 1 Italian & Art History = 1 Law = 1 Hispanic Culture = 1 Librarianship = 1 Medicine = 1 Global Health = 1 Management = 1 Natural Sciences = 2 Neuroscience = 2 Pharmacy = 3 Philosophy = 1 Psychology = 2 Russian and Italian = 1 Statistics = 1 Physics = 1 Urban Design = 1 Urban Planning = 2	Archaeology = 1 Biochemical Engineering = 1 Biochemistry = 2 Biotechnology = 2 Civil Engineering = 1 Classics = 1 Economics = 9 Engineering = 1 English Linguistics = 1 Financial Risk Management = 1 Fine Art = 1 History = 1 Human Science = 1 Languages = 1 Law = 5 Mathematics = 5 Medicine = 4 Management = 1 Natural Science = 1 Neuroscience = 2 Nutrition = 1 Pharmacy = 3 Physics and Biology = 1 Politics = 1 Psychology = 3 Security Services = 1 Speech Science = 2 Statistics = 1 Physics = 1 Urban Design = 1 Urban Planning = 2 Viking Studies = 1

Table 3.S1 Information on mean, median values and sample sizes for demographic data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-1.51	0.14	(-1.78, -1.24)	
Player type (strong)	0.62	0.28	(0.09, 1.17)	1.00
Game type (asymmetric)	0.17	0.21	(-0.23, 0.57)	1.00
Partner cooperated (yes)	1.94	0.16	(1.62, 2.27)	1.00
Partner punished (yes)	0.10	0.22	(-0.34, 0.54)	1.00
Player type x Game type	-0.96	0.42	(-1.78, -1.13)	1.00
Player type x Partner cooperated	0.31	0.33	(-0.36, 0.94)	1.00
Player type x Partner punished	-0.67	0.39	(-1.43, 0.09)	1.00
Game type x Partner cooperated	-0.62	0.34	(-1.28, 0.04)	1.00
Game type x Partner punished	0.80	0.46	(-0.09, 1.69)	0.70
Partner cooperated x Partner punished	-1.02	0.38	(-1.76, -0.28)	1.00
Player type x Game type x Partner cooperated	1.58	0.67	(0.25, 2.90)	1.00
Player type x Game type x Partner punished	0.79	0.77	(-0.72, 2.30)	0.27
Game type x Partner cooperated x Partner punished	1.10	0.74	(-0.35, 2.56)	0.52

Table 3.S2 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating players' responses to being punished for defecting in the previous round (player continued to defect = 0, player switched to cooperate = 1).

Chapter 4

Power Asymmetries and Punishment in a Prisoner's Dilemma with Variable Cooperative Investment

4.1 Note

This work is currently in preparation for submission. Nichola Raihani contributed to experimental design and discussion. Redouan Bshary contributed to discussion. Brian Wallace contributed to data collection and discussion. I designed the experiment, collected the data, analysed the data and wrote the paper.

4.2 Abstract

Recent work has found that in economic two-player experiments, players that invest in punishment finish with lower payoffs than those who abstain from punishing. These results have led researchers to question the effectiveness of punishment at promoting cooperation, especially when retaliation is possible. It has been suggested that these findings may stem from the unrealistic assumption that all players are equal in terms of power. However, a previous empirical study which incorporated power asymmetries into an iterated prisoner's dilemma (IPD) game failed to show that power asymmetries stabilise cooperation when punishment is possible. Instead, players' responses were conditional on their partner's behaviour in the previous round (i.e. players cooperated in response to their partner cooperating) and punishment did not yield any additional increase in tendency to cooperate. Nevertheless, this previous study only allowed an all-or-nothing – rather than a variable – cooperation investment. It is possible that power asymmetries increase the effectiveness of punishment from strong players only when players are able to vary their investment in cooperation. We tested this hypothesis using a modified IPD game which allowed players to vary their investment in cooperation in response to being punished. As in the previous study, punishment from strong players did not promote cooperation under any circumstances. Thus, it seems unlikely that human cooperation is promoted by either variable investment or power asymmetries in two-player games.

4.3 Introduction

Punishment involves paying a cost in order to inflict harm on cheats or defectors (Clutton-Brock & Parker 1995). Despite this cost, humans are often willing to invest in punishment in laboratory games involving both two players (Dreber et

al. 2011; Botelho et al. 2005; Camerer 2003) and multiple players (Bochet et al. 2006; Denant-Boemont et al. 2007; Egas & Riedl 2008; Fehr & Gächter 2002; Gürerk et al. 2006; Henrich et al. 2006; Nikiforakis & Normann 2008; Ostrom et al. 1992; Page et al. 2005; Rockenbach & Milinski 2006; Yamagishi 1986). Subjective pleasure from punishing others seems to be the proximate mechanism underlying such actions (de Quervain et al. 2004; Buckholtz et al. 2008). On a functional level, punishers may benefit from this investment if the target (or a bystander) behaves more cooperatively in future interactions (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012; dos Santos et al. 2013; Raihani & Bshary 2015).

However, empirical studies using two-player games have found that punishment reduces the payoffs of punishers (Dreber et al. 2008) and has no effect on the payoffs of their groups (Botelho et al. 2005; Dreber et al. 2008). Similarly, in multiplayer games punishment has been shown to have no effect (Bochet et al. 2006; Page et al. 2005) or even reduce (Egas & Riedl 2008; Fehr & Gächter 2002; Ostrom et al. 1992; Sefton et al. 2007) group payoffs in all but long -run encounters (Gächter et al. 2008). Moreover, when retaliation is possible, punishment reduces payoffs even in long-run encounters (Engelmann & Nikiforakis 2012). The threat of retaliation potentially increases the cost of punishment because punishers not only pay the cost of punishing itself but also any cost incurred if partners retaliate.

Previous studies have typically assumed that all players are equal in terms of power, meaning that all players can punish for the same cost and impose the same fine on targets (Botelho et al. 2005; Dreber et al. 2008; Bochet et al. 2006; Page et al. 2005; Egas & Riedl 2008; Fehr & Gächter 2002; Gächter et al. 2008; Ostrom et al. 1992; Sefton et al. 2007; Engelmann & Nikiforakis 2012). In reality, individuals are expected to vary in power, such that some players are able to inflict greater harm than their partners are able to reciprocate. When power asymmetries exist, it is expected that stronger players will punish weaker players but that weaker players will be unlikely to retaliate (Clutton-Brock & Parker 1995; Axelrod 1984; Raihani, Thornton, et al. 2012; Wang et al. 2010). This

prediction is borne out by data from the interspecific mutualism between cleaner fish (*Labroides dimidiatus*) and their reef-fish 'clients'. Cleaners provide a cleaning service to clients by removing skin ectoparasites (Grutter 1996). Although cleaner fish obtain nutrients from eating these ectoparasites, they prefer to eat the client's mucus, which constitutes 'cheating' (Grutter & Bshary 2003). If bitten, clients often terminate the interaction (Bshary & Grutter 2002). Cleaners sometimes work together in mixed sex pairs when cleaning a client. This creates a situation akin to a prisoner's dilemma because whilst only one cleaner can reap the benefits of eating the client's mucus, both share the cost of the interaction being terminated. A game theoretic analysis of this scenario demonstrated that for almost the entire parameter space, mutual defection is an evolutionary stable strategy (Bshary et al. 2008). Despite this, cleaner fish appear to have found a cooperative solution and pairs of interacting cleaner fish provide a better service to clients (more ectoparasite removal and less biting) than singletons (Bshary et al. 2008). The male fish are larger than the females and punish them by chasing them if they cheat (Raihani et al. 2010). Punished females behave more cooperatively in the next interaction with that male (Raihani et al. 2010). However, females never punish or retaliate against cheating males, apparently due to the size difference (Raihani, Grutter, et al. 2012; Raihani et al. 2010), suggesting that power asymmetries might stabilize cooperation in these mixed-sex interactions (Raihani, Pinto, et al. 2012; Raihani, Thornton, et al. 2012).

Power asymmetries may also stabilize cooperation in human social dilemmas by making punishment from strong individuals more effective at promoting cooperation than in symmetric games. Nevertheless, recent empirical work which has incorporated power asymmetries into economic games has failed to detect any positive effect of power asymmetries on the effectiveness of punishment for strong players. One such study explored the effects of power asymmetries in a public goods game (Nikiforakis et al. 2010). The authors found that asymmetries had no effect on punishment use, contributions to a public good or average payoffs. Although incurring punishment was shown to increase the contributions of low contributors, the authors did not test whether the effectiveness of punishment use was affected by power asymmetries. Moreover, in this study

retaliation was not possible because players were not informed which of their peers punished them. A more recent study explored the effects of power asymmetries in a two-player iterated prisoner's dilemma (IPD) game where retaliation was possible (Bone et al. 2015; Chapter 3 of this thesis). In this study, defecting players were more likely to cooperate in the next round if their partner cooperated, but punishment from the partner did not yield any additional benefits in either symmetric or asymmetric games. Moreover, counter to theoretical predictions, weak players were more likely to punish and retaliate in asymmetric games (i.e. against strong punishers) than in symmetric games. In fact, weak and strong players were equally likely to retaliate against strong partners (Bone et al. 2015).

One suggestion for why punishment from strong players failed to promote cooperation from weak partners in this setting is because cooperation was a binary decision: players could only choose between cooperate or defect. In such a setting, if the player's decision to cooperate (or defect) is conditioned on the partner's behaviour in the previous round, then there is little scope for punishment to have an additional positive effect on the behaviour of this individual. In fact, cooperation in real-life situations often involves a variable rather than an all-or-nothing investment (Frean 1996; Roberts & Sherratt 1998). For example, the cooperative allogrooming behavior exhibited in many animals (e.g. chacma baboons (*Papio cyanocephalus ursinus*)) may last from just a few seconds, up for several minutes (Barrett et al. 2000). Similarly, in the cleaner fish example, although defecting is a binary outcome (bite client / do not bite client) cooperative investment is variable (duration of time 'cooperating' by removing ectoparasites; Bshary et al. 2008). We therefore asked whether power asymmetries affected players' average cooperation levels as well as their tendency to (i) increase their cooperation investment and (ii) retaliate in response to being punished when cooperation was a variable rather than binary investment.

In order to address this question we used a modified version of the IPD game, with variable rather than binary investments. The game was structured such that increasing investments yielded mutual benefits but each player could gain a larger

benefit than the partner by choosing a slightly lower investment. Thus, the payoffs yielded the same incentive to defect as in the traditional prisoner's dilemma game (Table 4.1). Asymmetries were incorporated into the game by allowing strong players to interact with weak players. As in previous work (Nikiforakis et al. 2010; Bone et al. 2015) investing in punishment cost all players the same amount but strong players could inflict greater damage through punishing than weak players.

We predicted that players would be most likely to punish if their partner chose a lower cooperative investment than themselves (i.e. players would be more likely to punish defecting partners). Since previous work has found that in asymmetric games strong players were more likely to punish than weak players (Nikiforakis et al. 2010; Bone et al. 2015), we expected to replicate this pattern in this study. Based on theoretical and empirical insights (Clutton-Brock & Parker 1995; Raihani et al. 2010; Raihani, Grutter, et al. 2012), we predicted that being punished by a strong partner would induce weak players to increase their investment in cooperation in the next round, though we did not expect to find the same effect of punishment when strong players were punished by either weak or strong partners. Consequently, we envisaged that punishment from strong players would be more effective at promoting cooperation in asymmetric games than in symmetric games.

		Partner					
	Cooperation level	0	1	2	3	4	5
Player	0	(3, 3)	(6, 0)	(6, 0)	(6, 0)	(6, 0)	(6, 0)
	1	(0, 6)	(4, 4)	(7, 1)	(7, 1)	(7, 1)	(7, 1)
	2	(0, 6)	(1, 7)	(5, 5)	(8, 2)	(8, 2)	(8, 2)
	3	(0, 6)	(1, 7)	(2, 8)	(6, 6)	(9, 3)	(9, 3)
	4	(0, 6)	(1, 7)	(2, 8)	(3, 9)	(7, 7)	(10, 4)
	5	(0, 6)	(1, 7)	(2, 8)	(3, 9)	(4, 10)	(8, 8)

Table 4.1 Payoff matrix for players (player, partner) in step 1 of each round of the experiment. The player's cooperation level is given in the rows and their partner's cooperation level is given in the columns.

4.4 Methods

4.4.1 Experimental protocol

This research was approved by the University College London ethics board (project number 3720/001). All subjects remained anonymous so informed consent about the use of personal data was deemed unnecessary and was therefore waived by the University College London ethics board. The experiment took place over six sessions (one in May 2012, one in Nov 2012 and four in October 2014) in the experimental laboratory in the Department of Economics, University College London. The lab consists of twenty computers, which are visually partitioned. A total of 120 participants (71 women, 49 men, mean age ± se = 20.89 ± 0.20 years) were recruited from the student population to play a modified IPD game with a punishment option. Players interacted anonymously in pair-wise encounters by means of computer screens using the z-Tree (Fischbacher 2007) software. Each player played two games: one game with a partner of the same type as themselves (symmetric) and one game with a partner of a different type (asymmetric). The order in which players played symmetric and asymmetric

games was counter-balanced. All players were paid a £5.00 show-up fee and their final score was summed over both games and multiplied by £0.02 to determine additional earned income. Thus, one game unit corresponded to £0.02. To allow for negative incomes while maintaining the £5.00 show-up fee, all players began each game with 100 units (£2.00) to play with. The average payment per player was £19.34 and the average session length was 90 minutes. Prior to the experiment, each player was given written instructions about the game structure and required to answer ten comprehension questions to verify their understanding of the game (see supplementary materials for experimental instructions and questions). The average score from the comprehension questions was 88 %. Players were informed of the correct answers after the test.

The modified IPD game lasted 50 rounds. To avoid end effects (Rapoport & Dale 1966), players were told that each game would last between 20 and 100 rounds. Players' behaviour did not change abruptly towards the end of the game (Figure 4.S1), indicating that end effects were absent. Each round was split into two steps as follows:

Step 1: Both players simultaneously chose for how long they would like to cooperate with their partner. They could choose a time between zero and five seconds. For every second that both players cooperated they both got one unit. Whoever chose the shortest amount of time to cooperate for determined the duration of the interaction in that round and received a termination bonus of six units. If both players chose to interact for the same amount of time the interaction bonus was split into three units each. Hereafter, the amount of time a player chose to cooperate for will be called the players 'cooperation level'. After both players made their choice, they were shown the cooperation level they chose, whether their partner chose a higher, lower or equal cooperation level (but not the exact cooperation level chosen by their partner) and each player's payoffs from this step.

Step 2: Players were then given the option of whether or not to punish their partner (described below). At the end of step 2, players were shown their own and their partner's choice and payoff from step 2, as well as the cumulative payoffs for both players for that round and their own total payoff (summed over all rounds).

At the end of the first game, players were presented with the final scores and then randomly re-matched for the second game.

Players were randomly split into two types: weak and strong. Weak players punished with a 1 : 1 fee to fine ratio, meaning that if they chose to punish their partner it would cost them one unit and it would also cost their partner one unit. Strong players punished with a 1 : 6 fee to fine ratio, meaning that punishing their partner would cost them one unit but it would cost their partner six units. A 1 : 6 fee to fine ratio was chosen because the termination bonus was six units; thus, if players who chose to cooperate for a smaller amount of time than a strong partner were punished their payoff was lower than if they had chosen an equal or higher cooperation level than their partner.

To rule out the possibility that less powerful players were being coerced into a position where they would do better if they could avoid interacting with the aggressor altogether (e.g. (Nikiforakis et al. 2014)), players could choose to not participate (opt-out) in any round of the game. This option was presented in step one of each round of the game and meant that the current round was skipped and the next round then began as normal.

In order to avoid framing effects, neutral language was used. Player types “weak” and “strong” were replaced with “type 1” and “type 2”, “cooperate” was replaced with “interact” and “punish” and “don’t punish” were replaced by “option C” and “option D”. After both games had finished, all subjects were required to fill in a questionnaire to provide demographic information (see Supplementary materials for questions and demographic data).

4.4.2 Analyses

We used a series of generalised linear mixed models (GLMMs), to ask the following questions:

1. Did the mean cooperation level chosen by weak and strong players depend on whether they were in a symmetric or asymmetric game?

For each subject we first calculated the mean cooperation level chosen in each game they played. The mean cooperation levels were then set as the dependent term in a GLMM with the following explanatory terms: 'player type' (a 2-level factor with levels weak/strong), 'game type' (a 2-level factor with levels symmetric/asymmetric) and the two-way interaction 'player type x game type'.

2. Did the propensity of weak and strong players to punish their partner depend on whether they were in a symmetric or asymmetric game?

Punishment was coded as a binary response term (player did not punish = 0; player punished = 1) and set as the dependent variable in a GLMM, with the following explanatory terms: 'player type' (a 2-level factor with levels weak/strong), 'game type' (a 2-level factor with levels symmetric/asymmetric) and the two-way interaction 'player type x game type'. Instances of hypocritical or antisocial punishment (where the player punished the partner despite having chosen an equal or lower cooperation level than their partner, respectively) were not included in this model, leaving an N of 1735 rounds available for analysis.

3. Did punishing a cheating partner affect the punisher's total payoff?

The proportion of instances in which players punished a partner who chose a lower cooperation level than themselves was set as the dependent variable in a GLMM, with the following explanatory terms: 'player type' (a 2-level factor with levels weak/strong), 'game type' (a 2-level factor with levels symmetric/asymmetric) and the two-way interaction 'player type x game type'. Players whose partner never chose a lower cooperation level than themselves were not included in this analysis, leaving N of 198 players available for analysis.

4. Did being punished affect the likelihood that a player would increase their cooperation level in the next round in symmetric and asymmetric games?

Following Bone et al. (2015), we compared the likelihood that players increased their cooperation level in round *n+1* after having been punished (or not) in round *n*. Whether or not players increased their cooperation level in round *n+1* was coded as a binary response term (player didn't change or decreased cooperation level = 0; player increased cooperation level = 1) and set as the dependent variable in a GLMM, with the following explanatory terms: 'player type', 'game type', 'partner punished in round *n'* (a 2-level factor with levels no/yes), all two-way interactions and the three-way interaction. If the player or their partner opted out of either round n or round n+1 then both round n and round n+1 were excluded from the analysis. In addition, instances of antisocial or hypocritical punishment were not included in this model. Thus, data were restricted to instances where the player chose a lower cooperation level than their partner in round *n*, leaving an N of 1655 rounds available for analysis.

5. Did being punished increase the likelihood that a player would retaliate in the next round in symmetric and asymmetric games?

We classed a player as retaliating if they punished a partner who chose a higher cooperation level than themselves in round *n + 1* (i.e. 'antisocially' punished the partner), having been punished (restricted to justified punishment) by that partner in round *n*. Whether or not players punished their (cooperative) partner in round *n+1* was coded as a binary response term (player did not punish = 0; player punished = 1) and set as the dependent variable in a GLMM, with the following explanatory terms: 'player type' (a 2-level factor with levels weak/strong), 'game type' (a 2-level factor with levels symmetric/asymmetric), 'partner punished in round *n'* (a 2-level factor with levels no/yes), all two-way interactions and the three-way interaction. A positive effect of the term 'partner punished in round *n'* would indicate that players retaliated in response to being punished. If the player or their partner opted out of either round n or round n+1 then both round n and round n+1 were excluded from the analysis. Data were restricted to instances

where the player chose a lower cooperation level than their partner in round *n* and round *n+1,* leaving an N of 875 rounds available for analysis.

6. Did the propensity of weak and strong players to punish opt out depend on whether they were in a symmetric or asymmetric game?

The response term encoded whether the player opted out of the round (player didn't opt out = 0, player opted out = 1). The explanatory variables included in this model were: 'player type*'* (a 2-level factor with levels weak/strong), 'game type' (a 2-level factor with levels symmetric/asymmetric) and the two-way interaction 'player type x game type. All data were included in this analysis (N = 12000 rounds). This analysis showed that although opting out of rounds was generally rare, weak players in asymmetric games were more likely to opt out than players in other conditions (Table 4.2; Table 4.S2).

7. Did being punished increase the likelihood that a player would opt out in the next round in symmetric and asymmetric games?

Data were restricted to instances where the player had chosen a lower cooperation level than their partner in round *n* (i.e. the effect of antisocial or hypocritical punishment on the probability that the target opted out was not measured). Due to the small proportion of rounds in which strong players or weak players in symmetric games opted out (Table 4.2), we did not have the statistical power to test what factors affected opt out decisions for these players. Thus, data were restricted to weak players in asymmetric games (448 rounds were included in this analysis). Whether or not the player opted out of round *n+1* was coded as a binary response term (player didn't opt out = 0, player opted out = 1) and set as the dependent variable in a GLMM. The only explanatory variable included in this model was: 'partner punished in round *n'* (a 2-level factor with levels no/yes).

4.4.3 Statistical methods

Data were analysed using R version 2.15.2 (R Development Core Team 2011). Generalised linear mixed models (GLMMs) with Gaussian error structure and identity link function were used for analysis 1 and 3 and GLMMs with binomial error structure and logit link function were used for analyses 2, 4 & 5. GLMMs allow repeated measures to be fitted as random terms, thus controlling for their effects on the distribution of the data. For all models, player identity was included as a random term. Explanatory input variables were centred by subtracting their mean (Schielzeth 2010). After centring, continuous explanatory input variables were then standardized by dividing by 2 standard deviations.

4.5 Results

1. Did weak and strong players, respectively, chose different mean cooperation levels in asymmetric and symmetric games

Players often chose a non-zero cooperation level (mean proportion of rounds players chose non-zero cooperation level ± SE = 0.74 ± 0.03; Figure 4.1). Weak players chose higher cooperation levels in asymmetric games than in symmetric games (Table 4.2; Table 4.3; Figure 4.2). In contrast, strong players chose higher cooperation levels in symmetric games than in asymmetric games (Table 4.2; Table 4.3; Figure 4.2). Thus, both weak and strong players were most likely to cooperate if their partner was strong. In asymmetric games, weak and strong players chose equally high cooperation levels (Table 4.2; Table 4.3; Figure 4.2). In all conditions, mean cooperation levels increased slightly over the course of the game (Table 4.2; Figure 4.3). Cooperation levels appeared to increase more rapidly for strong players in symmetric games than for other players (Table 4.2; Figure 4.3).

2. Did the propensity of weak and strong players to punish their partner depend on whether they were in a symmetric or asymmetric game?

In general, players were most likely to punish if their partner chose a lower cooperation level than themselves ('justified punishment') than if their partner chose an equal cooperation level ('hypocritical punishment') or a higher cooperation level ('antisocial punishment'; Table 4.2). We investigated how the player's type (weak or strong) and game type (symmetric or asymmetric) affected their tendency to invest in justified punishment. Weak players were generally less likely to punish than strong players and, as expected, were more punitive in symmetric games than in asymmetric games (Table 4.2; Table 4.4; Figure 4.4). Strong players, on the other hand, were more likely to punish in asymmetric than symmetric games (Table 4.2; Table 4.4; Figure 4.4). Over the course of the game, use of justified punishment decreased for all but strong players in symmetric games; for these players justified punishment actually increased throughout the game (Table 4.2; Figure 4.3). In all conditions, use of hypocritical and antisocial punishment started low but decreased to even lower levels over the game (Table 4.2; Figure 4.3).

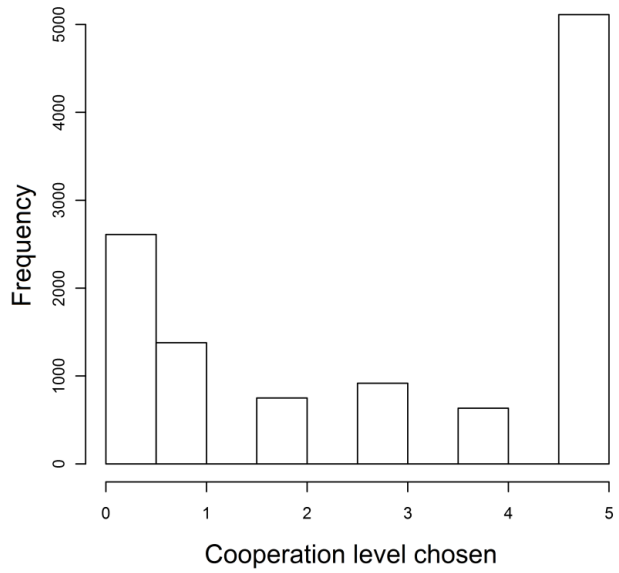


Figure 4.1 Histogram of cooperation levels chosen by players. Data exclude rounds where either player opted out.

	Weak players		Strong players	
	Symmetric	Asymmetric	Symmetric	Asymmetric
Cooperation level (whole game)	2.46 ± 0.23	2.91 ± 0.24	3.22 ± 0.22	2.94 ± 0.23
Cooperation level (last 10 rounds)	2.68 ± 0.09	3.19 ± 0.09	3.75 ± 0.08	3.17 ± 0.09
Justified punishment (whole game)	0.19 ± 0.04	0.14 ± 0.04	0.30 ± 0.05	0.41 ± 0.05
Justified punishment (last 10 rounds)	0.02 ± 0.02	0.11 ± 0.04	0.18 ± 0.06	0.56 ± 0.07
Hypocritical punishment (whole game)	0.01 ± 0.00	0.02 ± 0.00	0.05 ± 0.00	0.06 ± 0.01
Hypocritical punishment (last 10 rounds)	0.00 ± 0.00	0.02 ± 0.01	0.01 ± 0.00	0.03 ± 0.01
Antisocial punishment (whole game)	0.03 ± 0.01	0.07 ± 0.03	0.16 ± 0.04	0.17 ± 0.04
Antisocial punishment (last 10 rounds)	0.00 ± 0.00	0.02 ± 0.02	0.2 ± 0.06	0.03 ± 0.02
Opted out	0.01 ± 0.00	0.07 ± 0.02	0.02 ± 0.01	0.01 ± 0.00
Total payoff	376.09 ± 15.90	337.48 ± 19.66	311.19 ± 17.35	327.82 ± 16.89

Table 4.2 Summary data for mean cooperation level, mean proportion of instances where the player punished (justified / hypocritical / antisocial) and opted out and mean total payoffs (all means +/- SEM).

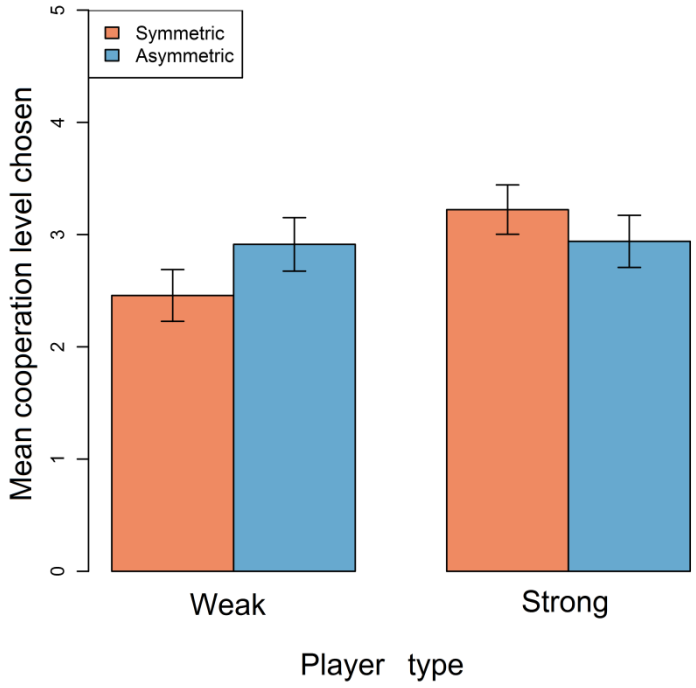


Figure 4.2 Barplot showing the mean (+/- SEM) cooperation levels chosen by weak and strong players in symmetric and asymmetric games. Data exclude rounds where either player opted out. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	2.88	0.14	(2.61, 3.15)	
Game type (asymmetric)	0.09	0.17	(-0.25, 0.42)	0.71
Player type (strong)	0.40	0.27	(-0.15, 0.94)	0.41
Game type x Player type	-0.74	0.34	(-1.41, -0.07)	0.41

Table 4.3 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected the mean cooperation level chosen the player in each game. Data were restricted to instances where the player chose a higher cooperation level than their partner in that round.

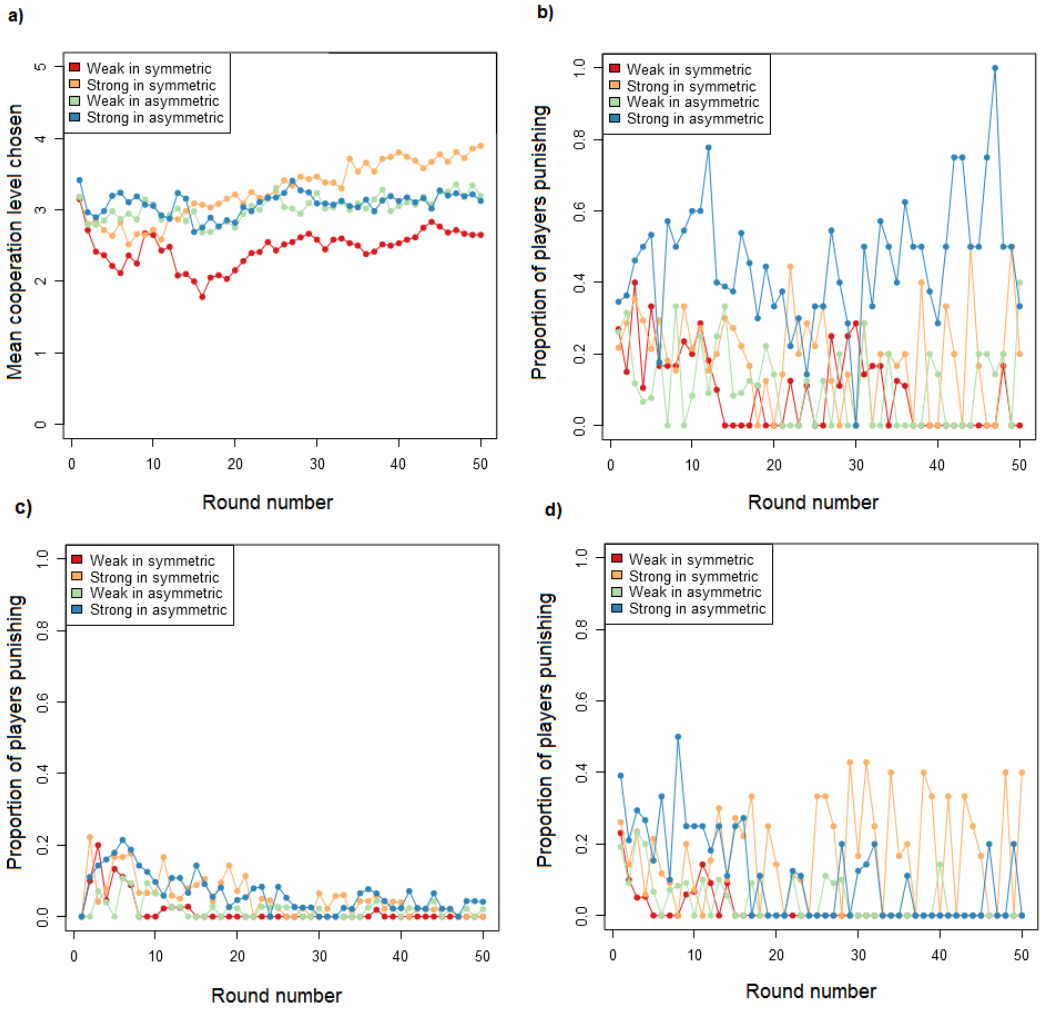


Figure 4.3 Scatter plot showing **a)** the mean cooperation level chosen; and the mean proportion of players that chose **b)** justified punishment **c)** hypocritical punishment and **d)** antisocial punishment in each round according to whether they were weak or strong and whether they were in a symmetric or asymmetric game. Rounds where either player opted out were excluded.

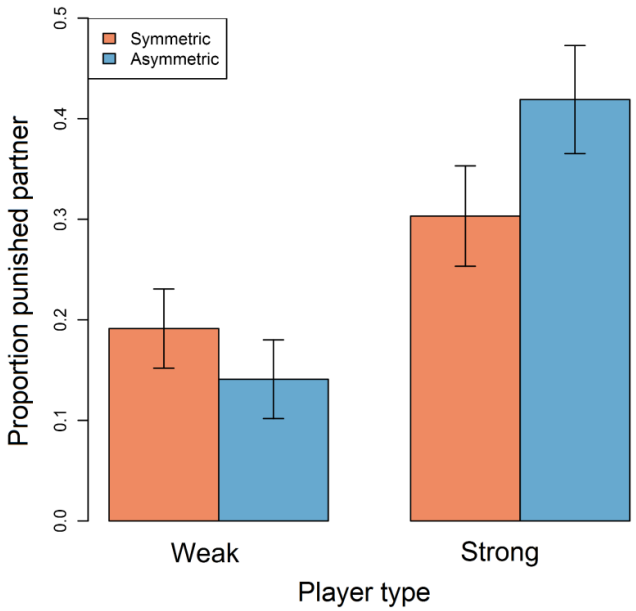


Figure 4.4 Barplot showing the mean (+/- SEM) proportion of instances in which weak and strong players punished their partner in symmetric and asymmetric games. Data were restricted to instances where the player chose a higher cooperation level than their partner and exclude rounds where either player opted out. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-1.84	0.22	(-2.31, -1.42)	
Game type (asymmetric)	0.09	0.19	(-0.28, 0.45)	1.00
Player type (strong)	1.70	0.43	(0.86, 2.59)	1.00
Game type x Player type	1.68	0.37	(0.96, 2.42)	1.00

Table 4.4 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected whether players punished their partners (player did not punish partner = 0; player punished partner = 1). Data were restricted to instances where the player chose a higher cooperation level than their partner in that round.

3. Did punishing a cheating partner affect the punisher's total payoff?

Justified punishment had no effect on the punisher's payoff, regardless of player type or game type (Table 4.5; Figure 4.5). Nevertheless, weak player generally received higher payoffs than strong players (Table 4.2; Table 4.5). Although it appears that both weak and strong players received lower payoffs when interacting with a strong partner, the confidence interval for the interaction term crossed zero meaning evidence for this effect is weak (Table 4.2; Table 4.5).

4. Did being punished affect the likelihood that a player would increase their cooperation level in the next round in symmetric and asymmetric games?

Justified punishment had no discernible effect on the target's tendency to increase their cooperation level in the next round, regardless of whether they were weak or strong or the game type (Table 4.6; Figure 4.6).

5. Did being punished increase the likelihood that a player would retaliate in the next round in symmetric and asymmetric games?

Weak and strong players both retaliated in response to justified punishment (i.e. they were more likely to punish a cooperative partner if this partner had punished them in the previous round than when the partner had not punished in the previous round; Table 4.7; Figure 4.7). Contrary to our predictions, neither the player's type nor game type had an effect on whether players retaliated against punitive partners (Table 4.7; Figure 4.7). Although strong players in symmetric games appeared to retaliate more frequently than players in other conditions (Figure 4.7), the 3-way interaction between the players type, the game type and whether or not the player was punished by their partner in the previous round was not a component of the top models, meaning evidence for this effect is weak.

6. Did the propensity of weak and strong players to punish opt out depend on whether they were in a symmetric or asymmetric game?

Although opting out of rounds was generally rare, weak players in asymmetric games were more likely to opt out than players in any other condition (Table 4.2; Table 4.8).

7. Did being punished increase the likelihood that a player would opt out in the next round in symmetric and asymmetric games?

We did not have the statistical power to test what factors affected opt out decisions for all but weak players in asymmetric games. However, weak players in asymmetric games were more likely to opt out if they were punished by their partner in the previous round (proportion opting out if partner didn't punish = 0.00 ± 0.0 versus if partner punished 0.09 ± 0.02; Table 4.9).

Parameter	Estimate	SE	Confidence Interval	Importance
Intercept	339.07	9.30	(320.72, 357.41)	
Player type (strong)	-37.88	18.60	(-74.56, -1.20)	1.00
Game type (asymmetric)	-13.80	15.74	(-44.84, 17.24)	0.56
Player type x Game type	54.06	31.31	(-7.70, 115.82)	0.33

Table 4.5. Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models for the response term encoding 'total payoff'.

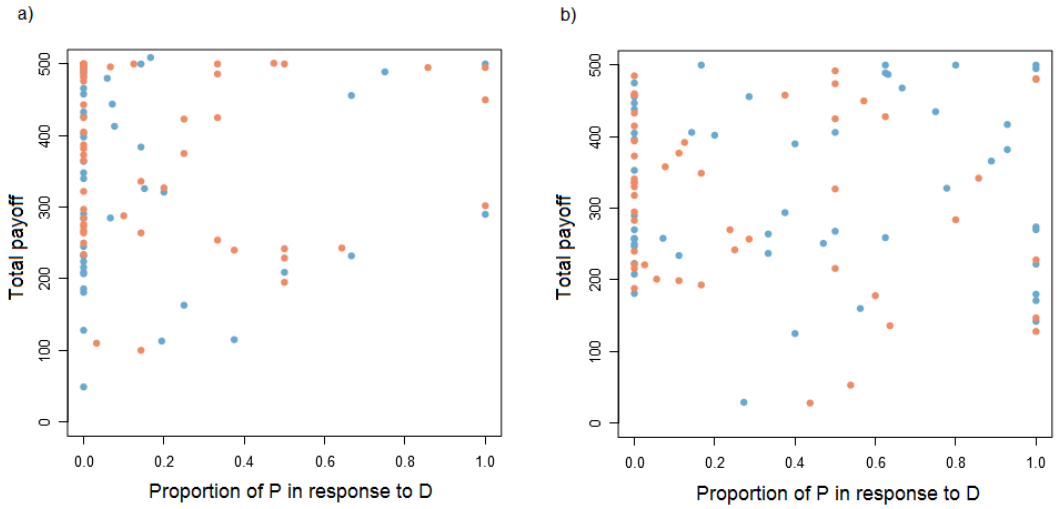


Figure 4.5. Scatter plots showing how the probability of punishing in response to defection affected total payoffs in symmetric and asymmetric games for (a) weak players and (b) strong players. Circles represent raw data points; red circles are symmetric games and blue circles are asymmetric games.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-0.09	0.14	(-0.37, 0.19)	
Partner punished in round *n* (yes)	-0.19	0.16	(-0.50, 0.12)	0.43
Player type (strong)	0.31	0.29	(-0.25, 0.88)	0.31
Game type (asymmetric)	0.10	0.13	(-0.17, 0.36)	0.22

Table 4.6 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected whether or not players increased their cooperation level in round *n+1* relative to round *n* (player didn't change or decreased cooperation level = 0; player increased cooperation level = 1). Data were restricted to instances where the player chose a lower cooperation level than their partner in round *n*.

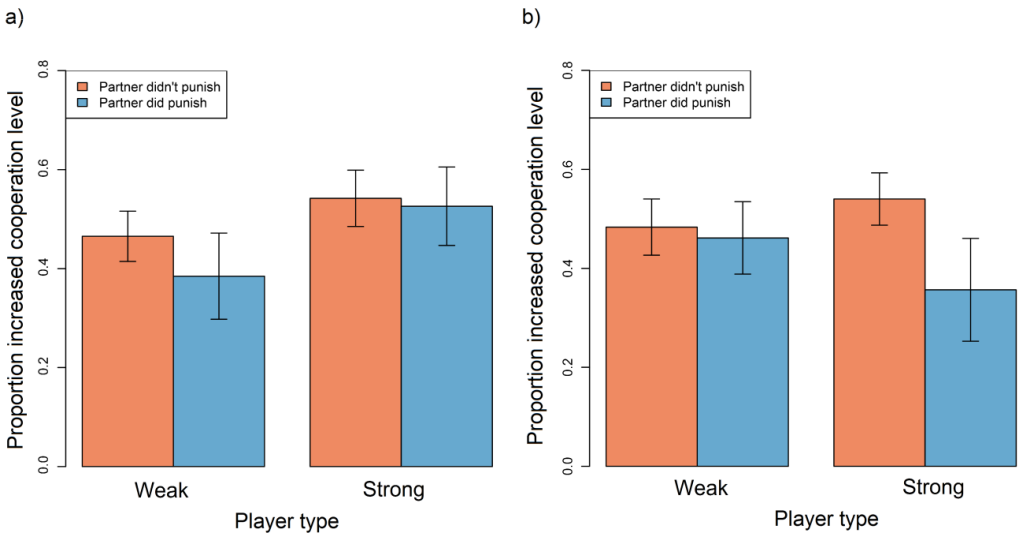


Figure 4.6 Barplot showing the mean proportion of instances in which weak and strong players in **a)** symmetric games and **b)** asymmetric games increased their cooperation level in round *n+1* relative to round *n*, according to whether or not they were punished by their partner in round *n*. Data were restricted to instances where the player chose a lower cooperation level than their partner in round *n* and exclude instances where either player opted out in round *n* or *n+1*. Error bars represent standard errors. Plots are generated from raw data.

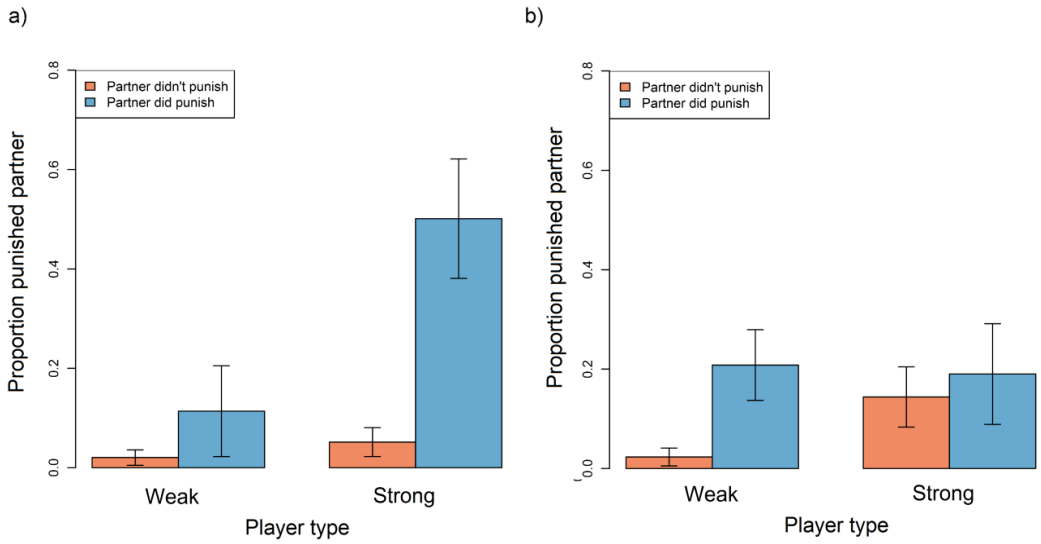


Figure 4.7 Barplot showing the mean proportion +/- SE of instances that weak and strong players in **a)** symmetric and **b)** asymmetric games punished a more cooperative partner in round *n+1*, according to whether they were punished by their partner in round *n*. Data were restricted to instances in which the player choose a lower cooperation level than their partner in round *n* and *n+1*. Rounds where either player opted out in round *n* or *n+1* were also excluded. Red bars represent antisocial punishment whereas blue bars can be interpreted as retaliation for punishment previously received. Plots are generated from raw data.

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-3.83	0.50	(-4.81, -2.86)	
Partner punished in round *n* (yes)	2.71	0.46	(1.81, 3.61)	1.00
Player type (strong)	1.75	0.69	(0.43, 3.06)	1.00
Game type (asymmetric)	0.41	0.41	(-3.85, 1.21)	0.44
Game type x Partner punished in round *n*	-1.13	0.84	(-2.77, 0.51)	0.21
Partner punished in round *n* x Player type	0.19	0.87	(-1.51, 1.89)	0.15

Table 4.7 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected whether a player punished a cooperative partner (player did not punish = 0; player punished = 1). Data were restricted to instances the player chose a lower cooperation level than their partner in round *n* and round *n+1*. The term 'partner punished in round n' describes whether the punishment can be interpreted as retaliation (i.e. player retaliating against a punitive partner) or antisocial punishment (i.e. player punishing a cooperative partner).

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-9.88	1.01	(-11.91, -7.94)	
Players type (strong)	0.18	0.93	(-1.66, 2.05)	1.00
Game type (asymmetric)	1.17	0.19	(0.79, 1.56)	1.00
Players type x Game type	-4.42	0.39	(-5.21, -3.69)	1.00

Table 4.8 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected whether the player opted out of rounds of the game (player did not opt out = 0; player opted out = 1).

Parameter	Effect size	SE	Confidence Interval	Importance
Intercept	-10.18	2.85	(-17.08, -5.86)	
Partner punished in round *n*	3.32	1.52	(0.93, 7.21)	1.00

Table 4.9 Effect sizes, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models investigating which factors affected whether the player opted out of rounds of the game (player didn't opt out of round $n+1$ = 0, player opted out of round $n+1$ = 1). Data are restricted to weak players in asymmetric games in instances where the player chose a lower cooperation level than their partner in round *n*.

4.6 Discussion

Animal research has been suggested that asymmetries in punishing power may stabilize cooperation in humans by making punishment from strong individuals more effective at promoting cooperation than in symmetric games (Raihani, Pinto, et al. 2012; Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012; Úbeda & Duéñez-Guzmán 2011). However, a previous empirical study which incorporated power asymmetries into an iterated prisoner's dilemma (IPD) game failed to find any positive effect of power asymmetries on cooperation (Bone et al. 2015). This failure was potentially due to decisions involving an all-or-nothing – rather than a variable – cooperation investment (Bone et al. 2015). We used a modified IPD game to test how power asymmetries affect the use of punishment and its effectiveness at promoting cooperation when both variable cooperation investment and retaliation were possible.

As expected, players were most likely to punish partners that chose a lower cooperation level than themselves and as observed in previous studies (Nikiforakis & Normann 2008; Ostrom et al. 1992; Falk et al. 2005; Egas & Riedl 2008; Nikiforakis et al. 2010; Bone et al. 2015), strong players were more likely than weak players to punish their partners. In addition, strong players were more likely to punish a weak than a strong partner. Weak players, on the other hand, were less likely to punish when faced with a strong partner. These findings support the prediction that punishment is most likely to operate down a dominance hierarchy (Clutton-Brock & Parker 1995; Axelrod 1984; Raihani, Thornton, et al. 2012; Wang et al. 2010; Bshary et al. 2008).

Other findings did not support our predictions based on animal research (Raihani et al. 2010). For example, we predicted that weak players would respond to punishment from strong partners with increased cooperation in the next round, whereas we did not expect such an effect when strong players were punished by either weak or strong partners. However, we found that in all conditions, punishment had no effect on a target's tendency to increase their cooperation level in the next round. Moreover, players' tendency to retaliate against punishers did not vary with game type for either strong or weak players. These findings are

consistent with previous work which incorporated power asymmetries into an IPD (Bone et al. 2015) and suggest that, in two-player prisoner's dilemma games at least, punishment might not be an effective strategy for motivating partners to cooperate.

Instead, the fact that punishment rarely induced partners to cooperate and often invoked retaliation supports the idea that conditionally cooperative strategies might often outperform punitive strategies in two-player games (Rand et al. 2009). Such conditionally cooperative strategies are expected to be less effective in multiplayer games where defection harms cooperative partners as well as defectors (Raihani, Thornton, et al. 2012). Punishment may therefore be more effective in multiplayer games than in two-player games (e.g. Przepiorka & Diekmann 2013).

Empirical studies using repeated public goods games have shown that players increase their cooperation levels if they know their peers are able to punish them (compared to a no punishment treatment) even if they are not informed whether or not they were punished after each round (Vyrastekova et al. 2008; Fudenberg & Pathak 2010). These studies suggest that the mere threat of punishment may deter cheating; regardless of whether punishment is actually observed (Vyrastekova et al. 2008; Fudenberg & Pathak 2010). In other words, knowing they face the possibility of being punished may lead individuals to increase their cooperation levels *a priori* to being punished. Although in our study players' tendency to increase their cooperation level in response to punishment was not affected by player type or game type, throughout the game, weak players generally chose higher cooperation levels in asymmetric games than in symmetric games. This finding suggests that the mere threat of punishment from a strong partner may have deterred weak players from defecting more effectively than the threat of being punished by a weak partner; even though actually incurring punishment did not change players' behaviour. If the threat of punishment sufficiently deters weak players in asymmetric games from cheating then strong players in asymmetric games would be required to punish less often (Cant 2011). This would reduce the cost associated with punishment for strong players in asymmetric

games; along with conveying benefits related to interacting with a more cooperative partner. It is possible that strong players would take the threat of being punished by a strong partner less seriously because they possess a credible threat of retaliation of their own (Cant 2011). If this were the case, then we would have expected that strong players would cooperate less than weak players when paired with a strong partner. However, strong players in symmetric games actually chose higher mean cooperation levels than weak players in asymmetric games, suggesting that the threat of being punished by a strong partner deterred cheating regardless of whether the player was weak or strong. Perhaps more puzzling is the finding that weak and strong players were equally cooperative in asymmetric games. It is possible that the high levels of cooperation exhibited by strong players in asymmetric games were a result of conditionally cooperative strategies, whereby strong players behaved cooperatively because they believed their weak partners would cooperate as well (Croson 2001; Keser & van Winden 2000; Fischbacher et al. 2001; Fischbacher & Gaechter 2006; Bone et al. 2015); rather than because they feared being punished. These findings are consistent with the Bone et al. (2015) study where it was found that while incurring punishment did not elicit cooperation from targets in the following round, players were generally more cooperative if their partner was strong. Although the implications of this finding were not discussed in the earlier study (Bone et al. 2015), together these studies suggest that the threat of punishment from a strong player may be sufficient to promote cooperation (Cant 2011), even if actual punishment has no effect. Further work is required to understand how the threat of punishment (even when never implemented) promotes cooperation when power asymmetries are present. Future work could incorporate power asymmetries into a set-up similar to that used by Vyrastekova et al. (2008) & Fudenberg & Pathak (2010), whereby players were not informed of whether or not they were punished by their peers after each round of the game (Vyrastekova et al. 2008; Fudenberg & Pathak 2010).

Although weak players were less likely to punish strong partners than weak partners, this effect was relatively small and weak players did still often punish in asymmetric games. The fact that weak players punished and retaliated against strong partners in these human experiments but in the cleaner fish system**,** female

fish never punish or retaliate against larger, dominant males (Bshary et al. 2008; Raihani et al. 2010; Raihani, Grutter, et al. 2012) may be associated with the different costs associated with provoking aggressive responses from a more dominant partner. For example, in this experiment and Bone et al. (2015) punishment (or retaliation) from a strong partner meant losing a known and relatively small amount of money. However, for female cleaner fish the cost of associated with provoking punishment (or retaliation) from a male fish is unknown and could potentially be fatal. In addition, for female fish, retaliating against a punitive male carries a risk of escalating aggression which was not possible in our game because opportunities to punish and the impact of punishment were fixed. The relatively small costs associated with being punished may also in part explain the ineffectiveness of punishment at promoting cooperation in both this study and Bone et al. (2015). Crucially, 'cheating' players received a higher payoff than their partner even if they were punished by a strong partner. Thus, if players are motivated by a desire to out-compete their partner as suggested in previous work (e.g. Fershtman et al. 2012; Houser & Xiao 2010), avoiding punishment may not have proved a sufficient incentive for players to behave more cooperatively. Future work should ask whether power asymmetries promote cooperation when the costs associated with retaliation are larger or have the possibility to escalate.

An alternative explanation for why weak players readily punished and retaliated against strong partners is that although we incorporated power asymmetries into the game these may have failed to translate into dominant and subordinate social roles in players' minds. This could stem from the use of neutral language in the game instructions given to participants. For example, although players were aware of the different payoff consequences of actions performed by the two player types, weak and strong players were referred to as 'Type 1' and 'Type 2' respectively. It is possible than these labels were not salient enough to elicit the behavioural responses we expected. This is a stark contrast to the famous Stanford prison experiment (Haney et al. 1973) where participants were randomly assigned the role of a prisoner or guard. In this experiment, effort was taken to make the situation as realistic as possible (e.g. guards were given sticks and uniforms and

prisoners role were arrested by the police department, deloused, forced to wear chains and prison garments). Under these conditions, within a short time both guards and prisoners settled into their new roles leading to extreme transformations of character (Haney et al. 1973). Other studies have shown that using loaded language like ‘bribe’ and ‘punish’ rather than neutral equivalents can produce significant changes in subjects' behaviour in economic games (e.g. Cameron et al. 2009). Although neutrally worded instructions have become a mainstream practice in behavioural experiments, it has been argued that it may be more useful to explore the effect of context rather than attempting the impossible goal of excluding it from experiments (Loomes 1999). We suggest that future work should explore how players behave in similar experiments when they are explicitly told that they are playing the role of a dominant or subordinate individual.

In this study punishing cheats had no discernible effect on punisher’s payoffs (as in Bochet, Page, & Putterman, 2006; Botelho, Harrison, Pinto, & Rutstrom, 2005; Page, Putterman, & Unel, 2005). Nevertheless, in the laboratory setting, individual payoffs are determined by the (largely) arbitrary costs and benefits associated with the options players are given in the game, as well as the number of rounds that players interact with one another (Gächter et al. 2008). In a real world setting, any fitness benefits of punishment must stem from its ability to promote cooperation from partners or bystanders. Thus, we suggest that since the net benefit of punishment in real-world settings must emerge from its ability to deter partners (or bystanders; dos Santos et al. 2013) from defecting, the effect of punishment on targets' behaviour, rather than the total payoffs accruing to the punisher, is more important for understanding the functional basis of punishment.

As observed in previous work (Bone el al. 2015), weak players in asymmetric games were considerably more likely to opt out than players in other conditions. In addition, weak players were more likely to opt out of rounds if they were previously punished by a strong partner than when they were not punished (see Supplementary materials for analysis). This suggests that weak players sometimes avoided further punishment from strong players by withdrawing from the game

rather than by increasing their cooperation level. In this study, players (especially weak players) opted out less often than in the previous Bone et al. (2015) study (players opted out of around 3% of rounds in this study vs. 11% of rounds in Bone et al. 2015). We suggest that players opted out less frequently in this study than in the previous study (Bone et al. 2015) because in the current study players could earn a positive payoff even if they chose a higher cooperation level than their partner (i.e. their partner defected). This was not the case in the previous study. Thus, unless they were the target of hypocritical or antisocial punishment, players in this study were always absolutely (if not relatively) better off participating rather than opting out.

To summarize, we found that in a variable investment IPD, power asymmetries did not make punishment from strong players more effective at promoting cooperation in comparison to symmetric games. In fact, punishment provoked retaliation, rather than cooperation in all conditions. This finding supports previous work which has suggested that in a two-player setting, conditional cooperation may sustain cooperation more effectively than punishment (Rand et al. 2009). We suggest that future research could explore the effect of power asymmetries when aggression can escalate. In addition, we propose that further work is required to understand how the threat of being punished influences players behaviour even when punishment is never implemented.

4.7 Supplementary materials

4.7.1 Experimental instructions

Welcome and thank you for participating in this experiment.

Please be quiet during the entire experiment. **Do not talk to your neighbours** and do not try to look at their screens. If you have any questions, please raise your hand. We will come to you and answer it privately.

This experiment is about decision-making. **You will be randomly assigned the role as either a "type 1" or a "type 2" player.** You will keep this role throughout the experiment. You will play a two-player game **twice**. Each game

will last a predetermined number of rounds **(between 20 and 100).**

In each game you **will be randomly matched with a different person in the room.** You will play with the same person for the duration of each game. Therefore, you will play with two people in total. In one game you will play with someone of the same type as yourself and in one game you will play someone of a different type to yourself. **You will remain anonymous throughout the experiment** and will be identified only by the name in the top left corner of your computer screen.

Before playing the games you will answer some questions about the experiment. The purpose of this is to make sure that everybody fully understands the rules of the experiment before we start so please make sure you read these instructions carefully.

Depending on your decisions and the decisions made by the other player in each of the two games you play, you will able to earn a considerable amount of money. The scores you receive in the game will be given in units **(1 unit= 2p).**

Everyone **will receive a show up fee of £5.** In addition, you will be given **an additional 100 units (£2.00) to play with at the start of each game**. Depending on the decisions made by you and the other player during each game you will gain or lose units from this initial amount. After the experiment has finished, your units will be converted to real money and you will be paid your earnings together with the show-up fee. In total this could be as much as **£29**.

Payment: Payment will happen after both games have been played.

The Game:

Each round will be split into two steps. In each step you will be asked to make a decision. The decisions you make will affect the income that you and the other player receive for that round. There is a time limit of **15 seconds** on each step (shown in the top right corner of the screen). It is important you make a choice within this time, if not **a default choice will be picked for you** and you will move on to the next step.

Each of these steps will now be explained to you in detail.

Step 1: You and the other player **simultaneously choose how long you would like to interact for (between 0 and 5 seconds).** To do this you will **select the button corresponding to the amount of time** you decide to interact for then **click OK.** Whoever chooses the smallest time determines the duration of the interaction in that round. For example, if you choose to interact for 3s and the other player chooses to interact for 5s the interaction will last for 3s. For **every second (s)** that you and the other player interact, **both players will get an income of +1 unit.**

If one player chooses to interact for less time than the other they will also receive a **termination bonus of +6 units**. **If both players choose to interact for the same amount of time** the termination bonus **of +6 units is split between the two players so that each player receives +3 units**.

You will also have the option of "**Do not participate in this round**". If either player chooses "Do not participate in this round" **both players will skip the second step and neither player will gain or lose any units in this entire round. The next round will begin as normal.**

Some examples are given below:

If both you and the other player choose to interact for 3s you will both receive +6 units (+3 for interacting for 3s, +3 split termination bonus = +6 total).

If you choose to interact for 2s and the **other player chooses to interact for 3s** you **will receive +8 units** (+2 for interacting for 2s, +6 termination bonus = +8 total).

If you choose to interact for 5s and the **other player chooses to interact for 0s** you will receive **0 units** 0 for interacting for 0s, 0 termination bonus = 0 total).

Step 2: If either player chose "Do not participate in this round" in step 1 then this step is skipped.

The decisions and scores from step one are presented to both players. You

must both then **decide** whether you would or would not like to reduce the other player's income at a cost to yourself by choosing between **options "A"** or **"B", respectively.**

Option "A" = reduce the other player's income at a cost of 1 unit to yourself.

Option "B" = do nothing (neither you nor the other player will gain or lose any points).

The number of units that the other player's income is reduced by when you choose option "A" varies depending upon what "type" of player you were assigned as at the start of the experiment.

Type 1 player: If you choose "A" then you will lose 1 unit, and the other player will lose 1 unit.

Type 2 player: if you choose "A" you will lose 1 unit, and the other player will lose 6 units.

After you have made a decision you and the other player are **presented with each other's choices and incomes from step 2** as well as your total income for the round. You will also be told your **total score** for the current game.

Your overall income in step 2 **is determined by the addition of the income from both your decision and the other player's decision**.

Your income for the round will be determined by the addition of your incomes from step 1 and step 2.

The total number of units that you have at the end of these games will determine how much money you have earned. Therefore, the additional money you and the other player each earn depends on the options you choose in each step of the game. However, **the final scores of the other players do not matter for your final earnings.**

If your **total score drops to 0 units**, you **will not be able to play** for the remaining rounds of the game of the **current game**.

At the end of both games, your total earnings will be computed. If you finish with a total score of **0 units over the two games**, you will walk away with just the £5 show up fee. If you have more than **0 units**, you will earn extra money at the exchange rate of 1 unit= 2p. The maximum extra amount that you can earn will be **£24**.

We will distribute a questionnaire at the end of the experiment that will ask some basic information about you.

Please take your time to read through the instructions again and if you have any questions raise your hand. In a few minutes we will begin the questionnaire followed by the games.

4.7.2 Comprehension questions

All participants were required to answer the following comprehension questions before playing the game. If they answered a question incorrectly they were shown the correct answer. The possible answers are shown in parentheses.

1. Each game will last a predetermined number of rounds, between what? (10-200 / 20-100 / 0-50)

2. How many games will you play in this experiment? (1 / 5 / 2)

3. If you finish the experiment with a total score of 500 units, how much money will you earn on top of the £5 show up fee (1 unit = 2p)? (£10 / £5 / £80)

4. If your total score drops to 0 or below in a game what will happen? (You will lose your show up fee / You will have to leave the room / You will not be able to play the remaining rounds of the current game)

5. If both you and the other player chose to interact for 3 seconds, how many units will you get in step 1? (1 / 3 / 6)

6. If you chose to interact for 4 seconds and the other person chose to interact for 5 seconds, how many units will you get in step 1? (10 / 1 / -6)

7. If you choose to interact for 2 seconds and the other player chooses to interact

for 1 second, how many units will you get in step 1? (2 / 7 / 1)

8. If you are a "type 1" player and choose option "A" in step 2, how many units will be deducted from the other players income? (1 / 3 / 5)

9. If both you and the other player are "type 2" players and you both choose option "A" how many units will you get in step 2? (1 / -7 / 2)

10. If both you and the other player are "type 1" players, you both choose to interact for 5 seconds in step 1 and both choose option "B" in step 2, what will your income be in that round? (8 / 2 / -5)

4.7.3 Demographic questions

After the game had finished, all participants were required to answer the following demographic questions. Responses to these demographic questions can be found in the supplementary data.

1. What is your gender?

2. How old are you?

3. What is your country of origin?

4. What is your subject of study?

4.7.4 Supplementary tables

Parameter	Weak	Strong
Age	Mean = 20.67 ± 0.31 Median = 20 IQR = 19 – 22 Range = 18 – 30	Mean = 21.12 ± 0.27 Median = 21 IQR = 20 – 22 Range = 18 - 28
Gender (n)	Females = 27 Males = 33	Females = 38 Males = 22
Country of Origin	Bangladesh = 1 Belarus = 1 Canada = 1 China = 4 Czech Republic = 1 Denmark = 1 France = 1 Greece = 1 Hungary = 2 India = 4 Indonesia = 1 Japan = 1 Jordon = 1 Malaysia = 6 Nepal = 1 Pakistan = 2 Peru = 1 Poland = 2 Russia = 2 Singapore = 5 Switzerland = 1 United Kingdom = 14 USA = 4 Zimbabwe = 1	Australia = 1 Azerbaijan = 1 China = 9 Czech Republic = 1 France = 1 Germany = 1 Greece = 2 Hong Kong = 5 Iceland = 1 India = 1 Indonesia = 1 Israel = 1 Italy = 1 Malaysia = 9 Poland = 1 Romania = 4 Russia = 1 Saudi Arabia = 1 Singapore = 4 Spain = 1 Taiwan = 2 Thailand = 1 United Kingdom = 8 USA = 1 Vietnam = 1
Subject Studied (n)	Anthropology = 1 Biochemistry = 2 Biology = 2 Biomedical Sciences = 2 Chemical Engineering = 2 Chemistry = 1 Economics = 5 Economics & Business = 1	Anthropology = 1 Archaeology = 1 Architecture = 1 Arts and Sciences = 2 Astrophysics = 1 Arts and Sciences = 1 Biology = 1 Brain and Mind = 1

Electronic Engineering = 2	Chemical Engineering = 1
Engineering = 4	Civil Engineering = 1
English = 1	Computer Science = 1
Fine art = 1	Digital Humanities = 1
French and German = 1	Economics = 7
French and Spanish = 1	Engineering = 1
Geography = 2	Environmental
History = 2	Fine art = 1
History of Art = 1	Geography = 1
History & Philosophy of	History = 1
Human Genetics = 1	Human Genetics = 1
Human Sciences = 1	Human sciences = 2
Language and Culture = 1	Infrastructure Investment
Translation theory = 1	Language and culture = 3
Management = 1	Law = 5
Mechanical Engineering = 1	Material Science = 1
Modern languages = 1	Mathematics = 2
Natural Science = 1	Medicine = 3
Pharmacogenetics = 1	Modern Languages = 1
Pharmacy = 2	Pharmacology = 1
Philosophy = 1	Pharmacy = 4
Philosophy and Economics = 1	Philosophy = 1
Physics = 4	Physics = 2
Political Science = 1	Psychology = 3
Politics = 2	Speech Sciences = 1
Psychology = 1	Statistics = 2
SSEES Politics, Security = 1	Statistics & Economics =
Systems Engineering = 1	Urban Planning = 1
Urban Planning = 1	Zoology = 1

Table 4.S1 Information on mean, median values and sample sizes for demographic data

Chapter 5

Human Punishment is Motivated by Both a Desire for Revenge and a Desire for Equality

5.1 Note

This work has been published as Bone JE, Raihani NJ (2015) doi: 10.1016/j.evolhumbehav.2015.02.002. Nichola Raihani contributed to experimental design and discussion. I designed the experiment, collected the data, analysed the data and wrote the paper.

5.2 Abstract

Humans willingly pay a cost to punish defecting partners in experimental games. However, the psychological motives underpinning punishment are unclear. Punishment could stem from the desire to reciprocally harm a cheat (i.e. revenge) which is arguably indicative of a deterrent function. Alternatively, punishment could be motivated by the desire to redress the balance between punisher and cheat. Such a desire for equality might be more indicative of a fitness-levelling function**.** We used a two player experimental game to disentangle these two possibilities. In this game, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced either advantageous inequality, equal outcomes or disadvantageous inequality. Players could punish stealing partners but some players had access to efficient punishment (1 : 3 fee to fine) whereas others could only use inefficient punishment (1 : 1). Players who had access to efficient punishment could reduce disadvantageous inequality by tailoring their investment in punishment whereas inefficient punishment did not change the relative payoffs of the individuals in the game but could be used to exact revenge. Players punished regardless of whether stealing created outcome inequality or whether punishment was capable of removing payoff differentials, suggesting that punishment was at least partly motivated by the desire to inflict reciprocal harm. However, in the efficient punishment condition, players' tendency to punish increased if stealing resulted in disadvantageous inequality and, when possible, punishers tailored their investment in punishment to create equal outcomes. Together these findings suggest that punishment is motivated by both a desire for revenge and a desire for equality. The implications of these findings are discussed

5.3 Introduction

Punishment typically involves paying a cost to harm individuals who harm or withhold benefits from the punisher (hereafter 'defectors', Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012; but see Irwin & Horne, 2013; Sylwester, Herrmann, & Bryson, 2013 for punishment aimed at helpful or cooperative individuals). Since punishment is costly to administer, both in terms of executing the punishment itself and in terms of the possibility of provoking retaliation from the target (Dreber et al. 2008; Herrmann et al. 2008; Janssen & Bushman 2008; Nikiforakis 2008), considerable effort has been expended in trying to understand the evolved function of punitive sentiments (McCullough et al. 2013; Price et al. 2002). Specifically, it has been argued that understanding the contexts that reliably motivate punishment can provide key insights into its likely evolved function (Price et al., 2002; but see Carlsmith, Darley, & Robinson, 2002). Two broad functional explanations have been proposed. First, it has been suggested that punitive sentiment could confer a selective advantage if punishment deters targets (or bystanders) from harming the punisher in future interactions (e.g. Dos Santos, Rankin, & Wedekind, 2011; Hilbe & Sigmund, 2010; McCullough et al., 2013). Under this hypothesis (hereafter the 'revenge' hypothesis), individuals should be motivated to reciprocally harm individuals that intentionally harm them, even if punishment cannot immediately equalize the payoffs between the defector and the punisher (Falk et al. 2005). However, evidence that punitive sentiments are sensitive to the risk of suffering a fitness disadvantage relative to defectors (Raihani & McAuliffe 2012a; Dawes et al. 2007) suggests an alternative explanation: that punishment primarily serves a fitness-levelling function, by reducing payoff differentials between defectors and punishers (Price et al. 2002). Under this fitness-levelling hypothesis, punishers are expected to be motivated primarily by the desire to equalize payoffs and any deterrent function of punishment would arise as a by-product. Here, we present an experiment to test whether punitive sentiment can best be explained in terms of desire for revenge or in terms of a desire to equalize payoffs in social interactions.

Interacting with a defector often reduces cooperators' payoffs and creates unequal outcomes. It can therefore be difficult to establish whether punishment of defectors is motivated by the disutility associated with receiving lower payoffs than a defector ('disadvantageous inequality aversion'; Fehr & Schmidt 1999) or simply a desire for revenge (Raihani & McAuliffe 2012b). A recent study attempted to disentangle these two possible motivations by asking whether, in the absence of disadvantageous inequality, experiencing losses was sufficient to motivate punishment (Raihani & McAuliffe 2012b). Raihani & McAuliffe (2012b) found that defection, in the form of stealing money from the victim, did not motivate punishment when stealing resulted in equal outcomes or advantageous inequality for the victim. However, stealing did motivate punishment when it resulted in disadvantageous inequality for the victim (Raihani & McAuliffe, 2012b). These findings raise the possibility that individuals use punishment to restore equality in social interactions. However, the alternative possibility, that punishment is simply related to the disutility associated with experiencing disadvantageous inequality and is not tailored to achieve equal outcomes, could not be ruled out because players in this game were not allowed to tailor their investment in punishment.

Alternative studies have also suggested that investment in punishment is aimed at producing equal outcomes in social interactions. For example, in Dawes et al. (2007) individuals were placed in groups of four and randomly allocated an endowment. Some players therefore started out richer than others in this game. Players were given the option to reduce (or increase) the income of others by purchasing negative (income-reducing) or positive (income-increasing) tokens and allocating these to other group members. In this setting, people allocated more negative tokens to the richest players and allocated more positive tokens to the poorest members of the group - suggesting that these behaviours were aimed at reducing outcome inequality. However, in this experiment, all four group members were able to purchase and allocate these tokens. Thus, it was impossible for players to predict how many tokens they would need to buy in order to achieve equal outcomes. Consequently, it is not possible to determine whether players adjusted investment in punitive behaviour in order to achieve specific outcomes.

Moreover, since initial payoff inequalities were exogenously determined rather than arising through some players defecting, the study could not test to what extent investment in income-reducing tokens was related to the target's behaviour, as opposed to the outcome itself. In other words, since cooperation and defection were not possible in this game, any revenge-based motives of punishment could not be measured.

A more recent study by Houser & Xiao (2010) showed that players who were treated unfairly most commonly chose to punish as severely as possible and thus create inequality in their own favour. Although this seems to be more suggestive of punishment as a form of revenge rather than a fitness-leveller, it is important to take into account that in this study the severity of punishment chosen was not constrained by cost. In reality, imposing a larger cost on another individual is likely to also impose a larger cost on the punisher (Raihani & McAuliffe 2012a). Since punishers have been shown to adjust their investment according to the costs associated with punishment (Anderson & Putterman 2006; Bone et al. 2014; Carpenter 2007; Nikiforakis & Normann 2008; Ostrom et al. 1992), this creates a potentially important trade-off between maximizing income and achieving the desired punishment outcome.

The fitness-levelling hypothesis predicts that individuals should only invest in punishment that is more costly to the target than to the punisher, and is therefore able to reduce any existing disadvantageous inequality. Nevertheless, empirical work has demonstrated that individuals are prepared to invest in punishment that is equally costly to the punisher and the target (Anderson & Putterman 2006; Carpenter 2007; Egas & Riedl 2008; Falk et al. 2005; Nikiforakis & Normann 2008) - or even more costly to the punisher (Anderson & Putterman 2006; Carpenter 2007; Egas & Riedl 2008) - and so is unable to re-establish equality. These findings suggest that punishers are not solely motived by a desire to remove fitness differentials and support the idea that punishers might instead be motivated by a desire for revenge against defecting partners. The predictions of the two hypotheses also differ with respect to whether the defection was performed intentionally or not. Specifically, the revenge hypothesis predicts that punishment

should be focused on those who impose harm intentionally and can therefore learn to avoid repeating the harmful behaviour in the future. Conversely, punishment aimed at removing fitness differentials should be less sensitive (or insensitive) to intentionality since the primary function is to reduce inequality rather than change the target's behaviour. Evidence from empirical studies provides some support for both hypotheses. Whilst several studies have shown that individuals will punish in response to unequal outcomes created at random or unintentionally (Cushman et al. 2009; Dawes et al. 2007; Falk et al. 2008; Houser & Xiao 2010; Kagel et al. 1996; Yu et al. 2014), individuals are significantly more likely to punish when unequal outcomes are created intentionally by the target (Falk et al. 2008; Houser & Xiao 2010; Kagel et al. 1996).

Based on past research it is therefore unclear whether punishment is motivated by a desire for revenge or by a desire to equalize payoffs. We aimed to answer this question by investigating whether victims of cheats adjusted their investment in punishment in order to restore equality using a modified version of the game used by (Raihani & McAuliffe 2012b). In the current study, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced advantageous inequality, equal outcomes or varying levels of disadvantageous inequality. Players could punish stealing partners, but while some players had access to efficient punishment (1 : 3 fee to fine), others could only use inefficient punishment (1 : 1 fee to fine). Players who had access to efficient punishment could achieve equal outcomes by tailoring their investment in punishment: more extreme outcome inequality could be alleviated by investing more into punishment. However, under the inefficient punishment condition, increasing investment in punishment did not reduce inequality.

Although we suggest that revenge may serve a deterrent function, in the anonymous one-shot setting of our game, there is no scope for punishment to change the behaviour of stealing partners (or bystanders). However, previous work has suggested that behaviour may be constrained by psychological mechanisms that evolved in the context of non-anonymous repeated interactions and that responses that are attuned to these conditions may be invoked even in

anonymous, one-shot settings (Ben-Ner & Putterman 2000; Burnham & Johnson 2005; Cosmides & Tooby 1989; Delton et al. 2011; Hagen & Hammerstein 2006; Hoffman et al. 1998; Johnson et al. 2003; Tooby et al. 2006). Thus, in our game a desire for revenge might reflect the desires of an evolved psychology that functions to deter cheats, even though this function is (due to the nature of the game) impossible to achieve. Nevertheless, we note that since deterrence is not the only possible function for this behaviour we use the word 'revenge' in a purely descriptive sense.

The revenge hypothesis predicts that punishment will be used in both the inefficient and the efficient punishment condition. Alternatively, if punishment is motivated by the desire to equalize outcomes, punishment should be used when it is efficient but not when it is inefficient. Moreover, players should use the amount of punishment that is required to equalize payoffs (Table 5.1); not more or less.

5.4 Methods

5.4.1 Experimental protocol

This research was approved by the University College London ethics board project number 3720/001. Data were collected in October 2013 and July - August 2014. We recruited 4912 subjects (2856 males, 1967 females, 89 unreported) for our experiment using the online labour market, Amazon Mechanical Turk (AMT; www.mturk.com; see General Methods for details). Subjects were all based in the USA.

Of the 4912 subjects, 2456 were assigned the role of player one (P1). The remaining subjects were allocated the role of player two (P2). P1 and P2 were both allocated one of five initial endowments (treatment A – E; Table 5.1). The game consisted of two stages. In the first stage P2 could choose to steal \$0.20 from P1 or do nothing. In the second stage, P1 was informed of P2's decision and could choose how many punishment points they wished to assign to P2. P1 experienced the same losses when P2 stole (\$0.20) in all five treatments. However, depending on the treatment this \$0.20 loss resulted in P1 experiencing either advantageous inequality (treatment A), equal payoffs (treatment B) or

disadvantageous inequality (treatments C – E; Table 5.1) relative to P2. All players were assigned to one of two punishment conditions at the start of the game: inefficient and efficient. In the inefficient punishment condition, each punishment point cost P1 $0.05 and reduced P2's earnings by $0.05 (fee to fine ratio = 1 : 1). In the efficient punishment condition, each punishment point cost P1 $0.05 and reduced P2's earnings by $0.15 (fee to fine ratio = 1 : 3). To prevent negative earnings, P1 could assign a maximum of four punishment points to P2.

P1 was assigned ex-post (Rand 2012) to one of two treatments in which either P2 stole or P2 didn't steal (Table 5.1). These treatments were allocated to players both in the inefficient and efficient punishment conditions, creating a total of 10 treatments for P2 and a total of 20 treatments for P1. All subjects that participated in the experiment received a $0.20 show-up payment on top of a bonus based on both their and their partner's decisions during the game.

5.4.2 Analyses

Data were analysed using R version 2.15.2 (R Development Core Team 2011). All comparisons used two-sided Fishers exact tests. First, we investigated whether experiencing losses or disadvantageous inequality had a greater effect on P1's decision to punish P2. We compared the proportion of P1s that chose a non-zero punishment investment when (i) P2 didn't steal (across all treatments), (ii) P2 stole but the stealing did not result in disadvantageous inequality for P1 (i.e. treatments A & B) and (iii) P2 stole resulting in disadvantageous inequality for P1 (i.e. treatments C – E). Separate analyses were conducted for players in the efficient and players in the inefficient punishment conditions (see Table 5.2 for comparisons and sample sizes).

Next, we investigated whether the inequality-removing punishment investment was picked more frequently than each of the other three possible punishment investments. Data were restricted to instances where P1s punished P2 for stealing in treatments where P2 stealing created disadvantageous inequality for P1 (treatments C – E) and when P1 had access to efficient punishment (see Table 5.3 for comparisons and sample sizes). We then asked whether these punitive players

were less willing to invest the amount required to create equality when doing so became progressively more expensive.

Finally, we investigated the possibility that players that chose the inequality-removing punishment investment may have done so because that amount of punishment was related to the disutility associated with the level of inequality experienced in that treatment, even when punishment was incapable of restoring equality (i.e. when punishment was inefficient). For this analysis we compared the proportion of players in the efficient punishment condition that chose the inequality-removing punishment investment versus the proportion of players in the inefficient punishment condition that chose that same punishment investment. Data were restricted to instances where P1s punished P2 for stealing in treatments where P2 stealing created disadvantageous inequality for P1 (treatments C – E).

As multiple comparisons were performed, sequential Benjamini - Hochberg adjusted p^{BH}-values (Benjamini & Hochberg, 1995; see also Waite & Campbell, 2006) are reported alongside uncorrected p-values. By controlling for the false discovery rate, Benjamini - Hochberg adjusted p-values balance the risk of incurring Type I errors with the risk of incurring Type II errors.

Treatment	Stage 1 payoff (P1 : P2)	P2 stole (Yes/No)	Stage 2 payoff (P1 : P2)	Outcome (from P1 point of view)	Efficient punishment investment required to create equal outcomes (cost to P1)
A	**$1.10 : $0.60**	**Yes**	**$0.90 : $0.80**	**Advantageous Inequality**	**NA**
		No	**$1.10 : $0.60**	Advantageous Inequality	NA
B	**$1.10 : $0.70**	**Yes**	**$0.90 : $0.90**	**Equal outcomes**	**NA**
		No	$1.10 : $0.70	Advantageous Inequality	NA
C	**$1.10 : $0.80**	**Yes**	**$0.90 : $1.00**	**Disadvantageous Inequality**	**$0.05**
		No	$1.10 : $0.80	Advantageous Inequality	NA
D	**$1.10 : $0.90**	**Yes**	**$0.90 : $1.10**	**Disadvantageous Inequality**	**$0.10**
		No	$1.10 : $0.90	Advantageous Inequality	NA
E	**$1.10 : $1.10**	**Yes**	**$0.90 : $1.30**	**Disadvantageous Inequality**	**$0.20**
		No	$1.10 : $1.10	Equal outcomes	NA

Table 5.1 The payoffs experienced by P1 and P2 at the beginning of Stage 1 and Stage 2 in treatments A - E. Stage 1 payoffs varied according to the treatment, while Stage 2 payoffs also depend on whether or not P2 stole. These payoffs are described in terms of the outcome (advantageous inequality, equal outcomes or disadvantageous inequality) from P1's point of view. Finally, we show the punishment investment that P1 was required to make to create equal outcomes when punishment was efficient.

5.5 Results

In both the efficient and the inefficient punishment condition, P1 was significantly more likely to punish a stealing than a non-stealing P2 (Fisher's exact test, see Table 5.2 for p-values; Figure 5.1). In the efficient punishment condition, the tendency to punish a stealing P2 was increased significantly when stealing resulted in disadvantageous inequality (proportion punishing non-stealing P2 ± SE = 0.04 ± 0.01; stealing P2, no disadvantageous inequality = 0.19 ± 0.03; stealing P2, disadvantageous inequality = 0.34 ± 0.02; Table 5.2; Figure 5.1). Although players in the inefficient punishment condition also appeared to be more likely to punish a stealing P2 when stealing resulted in disadvantageous inequality, this finding was non-significant (proportion punishing non-stealing P2 ± SE = 0.03 ± 0.01; stealing P2, no disadvantageous inequality = 0.06 ± 0.01; stealing P2, disadvantageous inequality = 0.1 ± 0.02; Table 5.2; Figure 5.1).

Punishment condition	**Comparison**	**P-value**	**P^{BH}-value**	**n**
Inefficient	P2 didn't steal vs. P2 stole no DI	0.039	0.047	849
	P2 didn't steal vs. P2 stole DI	<0.001	<0.001	975
	P2 stole no DI vs. P2 stole DI	0.073	0.073	618
Efficient	P2 didn't steal vs. P2 stole no DI	<0.001	<0.001	856
	P2 didn't steal vs. P2 stole DI	<0.001	<0.001	987
	P2 stole no DI vs. P2 stole DI	<0.001	<0.001	627

Table 5.2 The p-values generated by Fisher's exact tests (two-sided) comparing the proportion of P1 that chose a non-zero punishment investment when (i) P2 didn't steal; (ii) P2 stole but the stealing did not result in disadvantageous inequality for P1 ('P2 stole no DI'); and (iii) P2 stole resulting in disadvantageous inequality for P1 ('P2 stole DI'). Comparisons were made for players in both the inefficient and the efficient punishment condition. The fourth column reports Benjamini-Hochberg adjusted p^{BH}-values. The final column shows the sample size (n) for that comparison.

When P2 stealing created disadvantageous inequality for P1 (treatments C - E), if P1 had access to efficient punishment, P1 could equalize outcomes by punishing P2. The specific punishment investment that would create equal outcomes depended on the treatment (Table 5.1). In treatments C - E, when punishment was efficient, the punishment investment that created equal outcomes was chosen significantly more often than any other possible investment (Fisher's exact test, see Table 5.3 for p-values; Figure 5.2); and this punishment investment was chosen significantly more frequently in the efficient than the inefficient punishment condition (Fisher's exact test, p-value = 0.024; p^{BH}-value = 0.027; n = 167; Figure 5.2). Moreover, in the efficient punishment condition, a punishing P1 was equally likely to choose the punishment investment that created equal outcomes in all three treatments where P2 stealing created disadvantageous inequality for P1 (C – E; Fisher's exact test, p-value = 0.698; p^{BH}-value = 0.698; n = 130; Figure 5.2), indicating that players' attempts to equalize outcomes were largely insensitive to the cost associated with doing so. All significant findings reported above remained significant after p-values were adjusted according to the Benjamini–Hochberg procedure (Benjamini & Hochberg 1995).

When punishment was efficient, in both treatments where P2 stealing did not create disadvantageous inequality for P1 (treatments A & B), if P1 used punishment, they were most likely to choose the harshest punishment option available (Figure 5.2; see supplementary materials for details). Due to the small proportion of P1 who chose to punish P2 when P2 didn't steal or when punishment was inefficient, we did not have the statistical power to test which punishment investments were most popular in these scenarios (see supplementary materials for power analysis, conducted using GPower; Erdfelder, Faul, & Buchner, 1996).

Treatment	Comparison (proportion ± SE)	P-value	P^{BH}-value	n
C	\$0.05 (0.53 ± 0.08) vs. \$0.10 (0.16 ± 0.06)	<0.001	<0.001	43
	\$0.05 (0.53 ± 0.08) vs. \$0.15 (0.05 ± 0.03)	<0.001	<0.001	43
	\$0.05 (0.53 ± 0.08) vs. \$0.20 (0.26 ± 0.07)	0.015	0.018	43
D	\$0.10 (0.57 ± 0.07) vs. \$0.05 (0.13 ± 0.05)	<0.001	<0.001	47
	\$0.10 (0.57 ± 0.07) vs. \$0.15 (0.11 ± 0.05)	<0.001	<0.001	47
	\$0.10 (0.57 ± 0.07) vs. \$0.20 (0.19 ± 0.06)	<0.001	<0.001	47
E	\$0.20 (0.62 ± 0.08) vs. \$0.05 (0.12 ± 0.05)	<0.001	<0.001	40
	\$0.20 (0.62 ± 0.08) vs. \$0.10 (0.25 ± 0.07)	0.001	0.002	40
	\$0.20 (0.62 ± 0.08) vs. \$0.15 (0.00 ± 0.00)	<0.001	<0.001	40

Table 5.3 Data are restricted to players that punished a stealing partner and had access to efficient punishment. P values are generated from two-sided Fisher's exact tests. Benjamini-Hochberg adjusted p^{BH}-values are also presented to account for multiple comparisons. The final column shows the sample size (n) for that comparison.

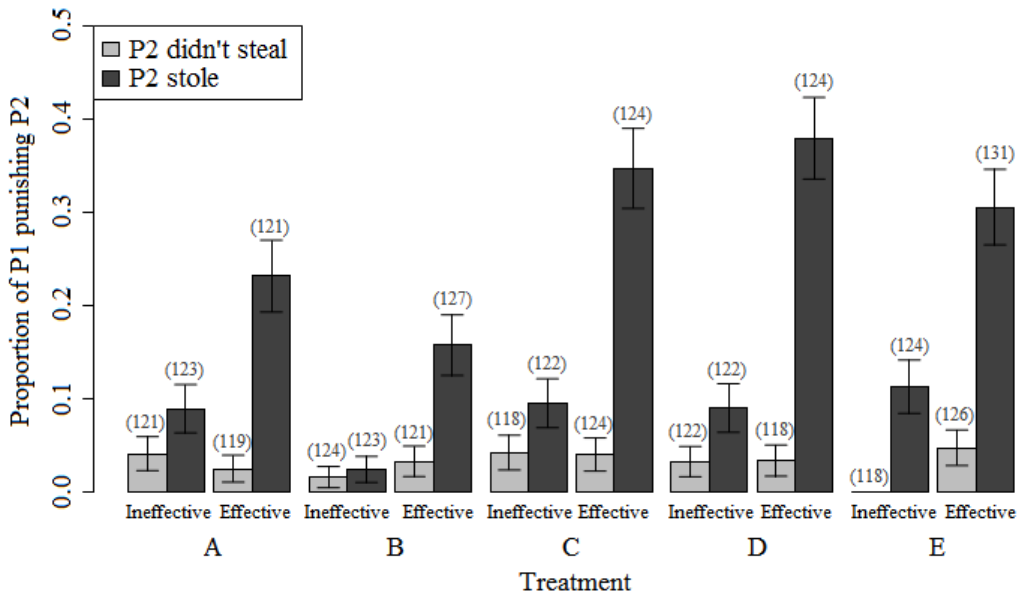


Figure 5.1 The proportion of P1 who punished P2 according to whether P2 stole (by taking $0.20 of P1's endowment), whether punishment was efficient (fee to fine ratio = 1 : 3) or inefficient (fee to fine ratio = 1 : 1) and the treatment. Initial endowments (P1 : P2) for treatment A were $1.10 : $0.60; in treatment B were $1.10 : $0.70; in treatment C were $1.10 : $0.80; in treatment D were $1.10 : $0.90 and in treatment E were $1.10 : $1.10. Thus, if P2 stole $0.20 from P1: in treatment A P1 experienced advantageous inequality ($0.90 : $0.80); in treatment B P1 experienced equal outcomes ($0.90 : $0.90) and in treatments C – E P1 experienced disadvantageous inequality ($0.90 : $1.00, $0.90 : $1.10 & $0.90 : $1.30, respectively). Sample sizes for each condition are indicated in parentheses. Light grey bars, P2 didn't steal; dark grey bars, P2 stole.

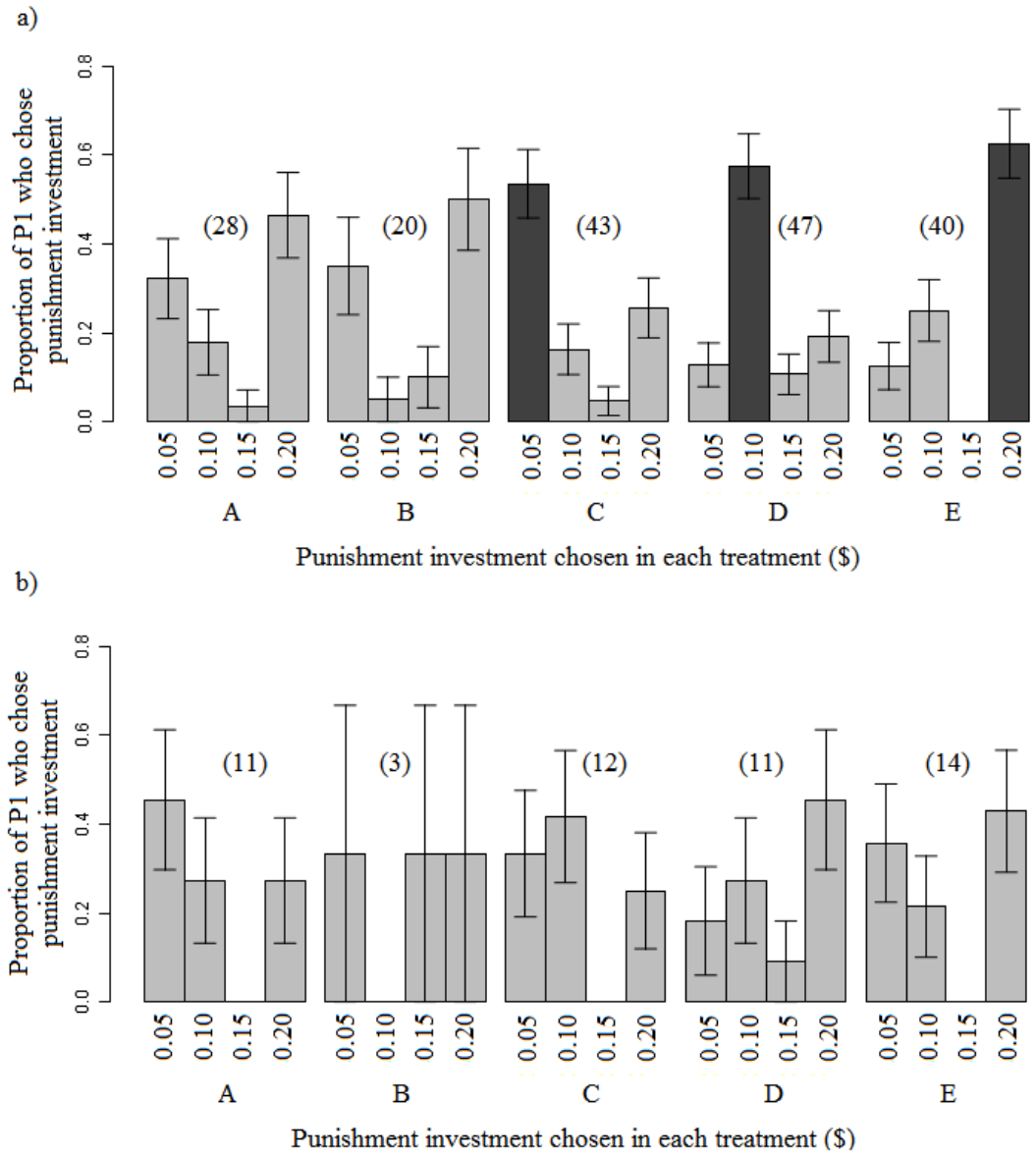


Figure 5.2 The proportion of punishment investments that were made in treatments A - E (given that P1 punished P2 for stealing) when punishment was **(a)** efficient and **(b)** inefficient. If punishment could create equal outcomes, the corresponding punishment investment is shown in dark grey for each treatment; all other punishment investments are shown in light grey. Sample sizes for each treatment are indicated in parentheses.

5.6 Discussion

In this study, P1 experienced the same losses when P2 stole ($0.20) in all five treatments. However, depending on the treatment this $0.20 loss resulted in P1 experiencing either advantageous inequality (treatment A), equal payoffs (treatment B) or disadvantageous inequality (treatment C - E) relative to P2. P2 stealing provoked P1 to punish even when stealing did not create disadvantageous inequality. Moreover, although relatively rare, P1 sometimes punished a stealing P2 even when punishment was inefficient and was thus unable to re-establish equality. Both these findings suggest that punishment was motivated at least in part by a desire for revenge against stealing partners. However, when punishment was efficient, P1 was more likely to punish if P2 stealing created disadvantageous inequality and, when given the option, P1 typically adjusted their investment in punishment to create equal outcomes. This suggests that although a desire for revenge was sometimes sufficient to motivate punishment, players were also sensitive to inequality and preferred punishment to result in equal outcomes.

Previous studies using three-player games have also shown that players will use apparently inefficient punishment (Egas & Riedl 2008; Falk et al. 2005). However, in these studies, it could be argued that, although inefficient punishment does not reduce inequality between players, it can reduce the standard deviation of the group's mean payoff and so may still be driven by egalitarian motives (Dawes et al. 2007). This, however, is not possible in two-player games like ours. In the current study, a willingness to pay for inefficient punishment therefore seems to reflect a desire for revenge (with an associated deterrent function, (McCullough et al. 2013) despite the fact that punishment occurred in an anonymous, one-shot setting where no deterrent function was possible. Although in our game punishment yielded no potential return on investment for punishers in terms of changing the partner's future behaviour), previous studies have proposed that the psychological mechanisms that underpin social behaviour (e.g. punishment) are likely to have evolved in a context where one-shot or anonymous interactions were rare (Delton et al. 2011; Fehr & Henrich 2003). It has been suggested that this evolved psychology may invoke responses that are attuned to these conditions

even in that are not adaptive in truly anonymous, one-shot lab settings (Fehr & Henrich 2003; Ben-Ner & Putterman 2000; Burnham & Johnson 2005; Cosmides & Tooby 1989; Delton et al. 2011; Hagen & Hammerstein 2006; Hoffman et al. 1998; Johnson et al. 2003; Tooby et al. 2006). Thus, it is possible that the use of inefficient punishment in our game was caused by the miss-firing of psychological mechanisms adapted to deter defecting partners from future defection, even when this function is (due to the nature of the game) impossible to achieve. Previous work has shown that when players are put under time pressure to make decisions in one-shot games, they are more likely behave cooperatively (Rand et al. 2014; Rand et al. 2012). Similarly, other studies also using one-shot games have shown that when players are given a cooling off period they are less likely to punish cheating partners (Grimm & Mengel 2011; Smith & Silberberg 2010; Sutter et al. 2003). These studies suggest that when players are the given time to consider their decisions they are more likely to respond in a way that maximizes their payoff in their current one-shot setting rather than rely on intuitions that may maximize payoffs over repeated encounters in the real world but not one-shot laboratory settings.

On the other hand, it could also be argued that the proximate mechanisms that underpin punishment may have evolved in a context where punishment was likely to have imposed larger costs on the target than the punisher (i.e. punishment was efficient) and was therefore capable of reducing the disadvantageous inequality experienced by victims of cheats. This line of reasoning would lead to the conclusion that a willingness to invest in inefficient punishment in our game could reflect the desires of an evolved psychology with the function of levelling fitness differentials (even though this function cannot be achieved in a context where punishment is inefficient). It is currently not clear what the most realistic fee to fine ratio is to use for punishment in laboratory settings in order to approximate the cost to impact ratio of punishment under real-world settings. Indeed, under real-world settings, the fee-to-fine ratio of punishment is likely to vary with relative dominance status of individuals (e.g. Bone, Wallace, Bshary, & Raihani, 2015; Raihani et al., 2012). Clearly, more studies of punishment in real-world settings are needed to establish how punishment use varies according to whether

interactions are repeated or not; and whether the fee to fine ratios currently used in the laboratory studies are ecologically valid.

Several players in this study used punishment to create advantageous inequality in their favour. For example, when players had access to efficient punishment and faced a stealing partner without also experiencing disadvantageous inequality (treatments A & B), punishing P1s typically chose the punishment investment that created the largest advantageous inequality for themselves. This finding is consistent with the idea that punishment is motivated by a desire for revenge, which might be 'sweeter' the more it harms the target (De Quervain et al. 2004); and is comparable to previous empirical findings which have shown that when the severity of punishment used was not constrained by cost, players often choose to punish as severely as possible and thus create inequality in their own favour (Abbink & Sadrieh 2009; Houser & Xiao 2010). This finding supports the idea that punishment sometimes stems from competitive motives, where players value being in a position of advantageous inequality because it emphasizes their relative social status (Fershtman et al. 2012; Houser & Xiao 2010).

Similar competitive motives have been inferred for the existence of 'antisocial' punishment (Herrmann et al. 2008; Sylwester et al. 2013; Raihani & Bshary 2015). As in several previous studies (e.g. Anderson & Putterman, 2006; Gächter, Herrmann, & Thöni, 2005; Gächter & Herrmann, 2009; Herrmann et al., 2008), we documented antisocial punishment (aimed at non-stealing partners) in this study. In this context, antisocial punishment cannot be explained by a desire for revenge or a desire to reduce disadvantageous inequality since P1 experienced neither losses nor disadvantageous inequality when P2 did not steal. It may be the case that antisocial punishment reflects competitive motives (Sylwester et al. 2013; Prediger et al. 2014; Raihani & Bshary 2015), though if this were the case we would have expected that players would use have used antisocial punishment in the efficient but not in the inefficient punishment condition, as previously documented (Falk et al. 2005). In contrast to this prediction, we found that players were equally likely to punish antisocially regardless of the punishment condition. It is possible that antisocial punishment in this study simply reflects execution

errors or misperceiving the game. With the current dataset we are unable to determine the causes of antisocial punishment but this remains an exciting avenue for future research.

Although many of our results support the idea that punishment was motivated primarily by a desire for revenge, we report two findings that support the hypothesis that punishment is motivated by a desire for equality (with an associated fitness-levelling function (Price et al., 2002). First, as in Raihani & McAuliffe (2012b), we found that in the efficient punishment condition players' tendency to punish a stealing partner was increased if stealing resulted in resulted in disadvantageous inequality. Second, when given the option, players typically tailored their investment in efficient punishment to remove disadvantageous inequality and were seemingly insensitive to the cost associated with achieving this outcome. Moreover, the punishment investment that created equal outcomes was chosen much more frequently in the efficient than the inefficient punishment condition, indicating that players were attempting to create equal outcomes rather than increasing punishment investment in response to frustration at experiencing increasingly disadvantageous outcomes.

Together our findings suggest that punishment is motivated by both a desire for revenge and a desire for equality. Indeed, these possibilities are not mutually exclusive. Furthermore, it might be the case that punishment which results in equality may be most likely to serve a deterrent function, if such punishment is perceived to be 'fair' and consequently more effective at changing the target's behaviour. This prediction is based on previous studies, where colleagues have suggested that 'morally legitimate' punishment is most likely to successfully deter future defection (Fehr & Rockenbach 2003; Houser & Xiao 2010). Fehr & Rockenbach (2003) suggest that punishment may be perceived as being morally illegitimate if it is associated with selfish or greedy (rather than altruistic) intentions. Punishment that creates advantageous inequality in the punisher's favour might be interpreted as a competitive act (Raihani & Bshary 2015) and therefore perceived as morally illegitimate. Punishment that creates advantageous inequality in favour of the punisher might therefore be unlikely to deter further

defection (Bone et al. 2015; Fehr & Rockenbach 2003; Xiao 2013) and may even provoke retaliation from the target (Bone et al., 2015). Whilst we stress that this explanation is speculative it offers promising avenues for further studies to explore the scenarios that motivate punishment.

We note that the current findings appear to contradict the results of Raihani & McAuliffe (2012b), who showed, using a similar experimental setup, that P1 only punished P2 where P2 stealing resulted in disadvantageous inequality. In contrast, in the current study, we found that players punished stealing partners even when stealing did not create disadvantageous inequality. We believe it is unlikely that the different costs of punishment used in the two studies ($0.05 in this study; $0.10 in the previous study) are responsible for these conflicting results (see supplementary materials for supporting analysis). However, it is possible that other subtle differences between our experimental setups may be responsible, specifically differences in the endowments initially given to P1 or differences in the demographic sample across the studies. In Raihani & McAuliffe's (2012b) experiment, the losses experienced by P1 as a result of P2 stealing ($0.20) were the same as in this experiment. However, the initial endowment of P1 was different: in Raihani & McAuliffe (2012b), P1 began the game with $0.70 and was left with $0.50 if P2 stole, whereas in this study, P1 began the game with $1.10 and was left with $0.90 if P2 stole. It has been shown that people pay most attention to the left-most digits when judging differences in the magnitude of numbers; a phenomenon known as the left-digit anchoring effect (Dehaene et al. 1990; Hinrichs et al. 1981; Monroe & Lee 1999; Thomas & Morwitz 2005). For example, an experimental study showed that a reduction of one cent affected the perceived magnitude of a price when the left digit changed ($3.00 to $2.99) but not when the left digit was unchanged ($3.20 to $3.19) (Thomas & Morwitz 2005). Thus, a reduction from $1.10 to $0.90 (as P1 experienced in this study) may be perceived as a greater loss than a reduction from $0.70 to $0.50 (as P1 experienced in Raihani & McAuliffe (2012b). If players perceived greater losses in this study than in Raihani & McAuliffe (2012b), this may explain why a $0.20 loss which did not result in disadvantageous inequality motivated P1 to punish P2 in this study but not in the earlier Raihani & McAuliffe (2012b) experiment. In

other words, in the absence of unequal outcomes, the loss experienced by P1 as a result of P2 stealing in Raihani & McAuliffe (2012b) may have been perceived as too small to motivate punishment.

Alternatively, the discrepancy between the results of the current study and Raihani & McAuliffe (2012b) may be explained by differences in demographic sampling between the two studies. Data for both studies were collected via the online labour market, Amazon Mechanical Turk, where the vast majority of workers hail from either the USA or India (Ross et al. 2010). In the Raihani & McAuliffe (2012b) study, participants were recruited from both countries and the analysis did not control for the possible cross-cultural differences in subjects' behaviour. However, in this study we restricted participation to subjects based in the USA. Previous studies have demonstrated cross-cultural differences in the propensity of subjects to punish both defectors (Henrich et al. 2006; Marlowe & Berbesque 2008) and cooperative individuals (Ellingsen et al. 2012; Gächter & Herrmann 2009; Herrmann et al. 2008). Thus, differences in the way that subjects from India versus the US behave in economic games, particularly with respect to punishment, may explain the different results we saw across the two studies. Future work will explore how cultural differences between players affect punishment strategies.

To summarize, we investigated whether punishment was motivated by a desire for revenge or a desire for equality and found support for both of these hypotheses. Players used punishment regardless of whether stealing created outcome inequality or whether punishment was capable at removing payoff differentials. This supports the hypothesis that punishment is motivated by revenge. However, players were more likely to punish if stealing resulted in disadvantageous inequality for the punisher and, when possible, typically tailored their investment in punishment to create equal outcomes. This supports the hypothesis that punishment is motivated by a desire to equalize payoffs. Since these hypotheses are not mutually exclusive we suggest that both a desire for revenge and a desire for equality are likely to play an important role in motivating punishment decisions. Future work should explore how the efficacy of punishment is related to its perceived moral legitimacy, and whether players are sensitive to this when

tailoring investment in punishment. We also suggest that more work is needed to understand what motivates antisocial punishment, how intuitions guide punishment decisions and in what ways cultural variation between players influences punishment strategies.

5.7 Supplementary materials

5.7.1 Demographic information

All participants were asked to answer the following demographic questions in Amazon Mechanical Turk (AMT). The options given are in parentheses:

What is your gender? (male / female)

What is your age?

Which of the following best describes your highest achieved education level? (Some High School / High School Graduate / Some College no degree / Associates degree / Bachelor's degree / Graduate degree)

What is the total annual income of your household? (Less than $12,500 / $12,500 – $24,999 / $25,000 - $37,499 / $37,500 - $49,999 / $50,000 - $62,499 / $62,500 - $74,999 / $75,000 - $87,499 / $87,500 - $99,999 / $100,000 or more)

Parameter	Individuals allocated to role of P1 (n = 2456)
Age	Mean = 28.76 ± 0.17
	Median = 27
	IQR = 23 – 32
	Range = 14 – 72
	Undisclosed = 39
Education level (n)	Some High School = 24
	High School Graduate = 237
	Some College, no degree = 873
	Associates Degree = 231
	Bachelor's Degree = 826
	Graduate Degree = 218
	Undisclosed = 47
Gender (n)	Females = 992
	Males = 1424
Annual income (n)	Less than $12,500 = 324
	$12,500 - $24,999 = 401
	$25,000 - $37,499 = 454
	$37,500 - $49,999 = 303
	$50,000 - $62,499 = 267
	$62,500 - $74,999 = 187
	$75,000 - $87,499 = 151
	$87,500 - $99,999 = 113
	$100,00 or more = 218
	Undisclosed = 38

Table 5.S1 Demographic information on age, education, gender and annual income levels for individuals allocated the role of P1.

5.7.2 Game instructions

Having completed the demographic questions above, participants were redirected to an external survey website (https://opinio.ucl.ac.uk) to take part in the experiment. Below is a text transcription of the game instructions received by players, including comprehension questions. The example given is for a player 1 in treatment A with efficient punishment; however, the general procedure is similar for all treatments and punishment conditions.

Screen 1. Please enter your Worker ID. This is needed to ensure you get your bonus. If you don't know your Worker ID you can find it out by opening the following page in a new window: https://www.mturk.com/mturk/dashboard.

Screen 2. **GAME INSTRUCTIONS** You are player 1. You have been allocated a bonus of $1.10. Player 2 has been allocated a bonus of $0.60. Your worker ID and player 2's worker ID will remain anonymous.

Screen 3. The game will be split into two stages: Stage one. Player 2 will choose between taking $0.20 of your bonus or doing nothing. If player 2 chooses to take $0.20 of your bonus, it will be added to player 2's own bonus. You will see player 2's decision. Stage two. You may pay a cost to reduce player 2's bonus or do nothing. Each time you reduce players 2's bonus it will cost you $0.05 and will reduce player 2's bonus by $0.15. You may choose to reduce player 2's bonus up to 4 times

Screen 4. Please answer these questions correctly to ensure your HIT is accepted. A. How much bonus have you been allocated? (You have been allocated a bonus of $1.10 / You have been allocated a bonus of $0.60).

Screen 5. B. How much bonus has player 2 been allocated? Player 2 has been allocated a bonus of $0.60 / Player 2 has been allocated a bonus of $1.10.

Screen 6. C. In stage one, will player 2 have the opportunity to take $0.20 of your bonus to add to their own bonus? Yes / No.

Screen 6. D. In stage two, you will have the opportunity to pay a cost to reduce player 2's bonus. You may choose to reduce player 2's bonus up to 4 times. How much will it cost you and by how much will it reduce player 2's bonus each time? It will cost you $1.00 and will reduce player 2's bonus by $0.60 each time / It will cost you $0.05 and will reduce player 2's bonus by $0.15 each time.

Screen 7. Well done - you got all the questions right! Ready to play the game? Yes.

Screen 8. **THE GAME** Stage 1. Player 2 could choose to take $0.20 of your bonus or do nothing. Player 2 chose to take $0.20 of your bonus. You initially had $1.10. You now have $0.90. Player 2 initially had $0.60. Player 2 now has $0.80.

Screen 9. Stage 2. You currently have $0.90 bonus. Player 2 currently has $0.80 bonus. You may pay to reduce player 2's bonus. Each time you reduce player 's bonus it will cost you $0.05 and will reduce player 2's bonus by $0.15. If you reduce player 2's bonus: 1 time - you will get $0.85 and player 2 will get $0.65 bonus. 2 times - you will get $0.80 and player 2 will get $0.50 bonus. 3 times - you will get $0.75 and player 2 will get $0.35 bonus. 4 times - you will get $0.70 and player 2 will get $0.20 bonus. If you don't pay to reduce player 2's bonus: You will get $0.90 bonus. Player 2 will get $0.80 bonus. How many times would you like to reduce Player 2's bonus? Don't pay to reduce player 2's bonus / 1 / 2 / 3 / 4.

Screen 10. That's the end of the game. The mystery word is 'cherry'. Please return to the HIT and enter the word 'cherry' in the box before submitting your HIT. Thanks for playing.

5.7.3 Matching of subjects

As described in the General Methods section it was not always possible to match players with a unique partner in the same treatment/punishment condition as them self. In this experiment 85.71 % of players (n = 4210) were matched with a unique partner in the same treatment and punishment condition as them self, 5.57 % of players (n = 274) were matched with a unique partner in the same

punishment condition but a different treatment and 0.16 % (n = 8) of players were matched with a unique partner in neither the same punishment condition nor the same treatment as them self. This left 110 P2s who did not steal and 110 P1s who were told that their partner did steal. These players (8.55 % of players) could not be uniquely matched with a partner and so were matched with a partner in the same treatment and punishment condition as them self but who had already been matched with another player (Figure 5.S1).

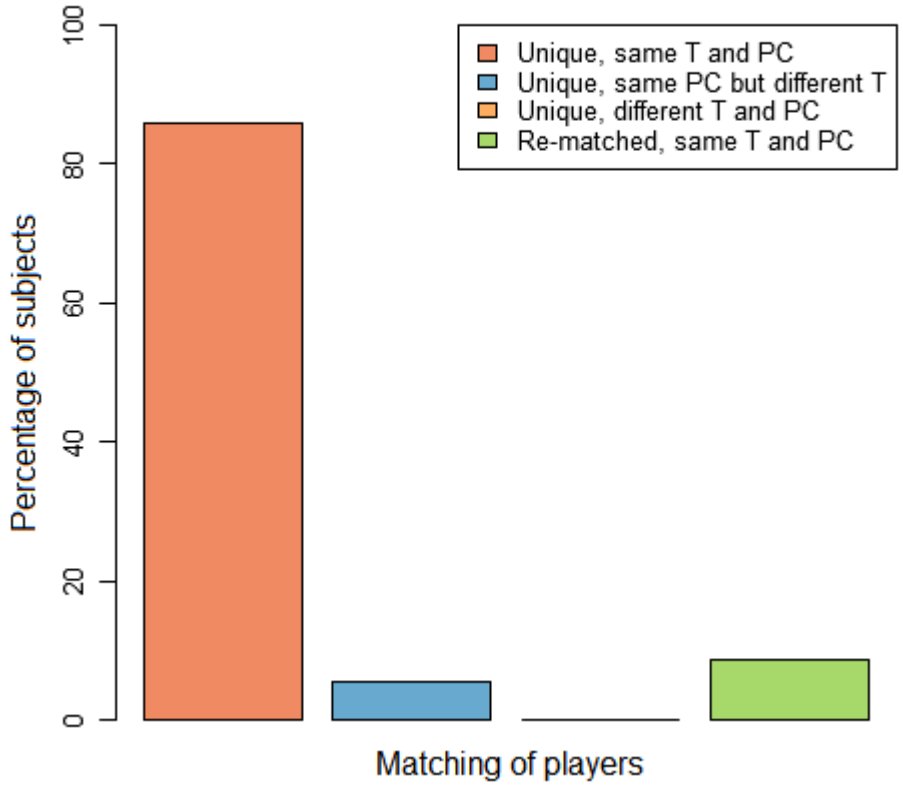


Figure 5.S1 The percentage of subjects that were matched with a unique partner in the same treatment and punishment condition (Unique, same T and PC; Red bar), a unique partner in the same punishment condition but a different treatment (Unique, same PC but different T; Blue bar), a unique partner in a different treatment and a different punishment condition (Unique, different T and PC; Orange bar) or a partner in the same treatment and punishment condition but who has already been matched with another player (Re-matched, same T and PC Green bar).

5.7.4 Power analysis

We used Gpower (Erdfelder, Faul, & Buchner, 1996) to conduct an *a priori* power analysis to assess whether our sample of P1s that punished was large enough to test whether P1 used the punishment option that would remove inequality in each treatment (C-E). For the power analysis, alpha was set at 0.05 (i.e. we did not account for any p-value adjustment to control for multiple comparisons), two tailed and power (1 - β) was set at 0.80.

The power analysis indicated that a sample size of 18 punishers would be required to detect a significant difference between a proportion of 0.70 (higher than observed in any treatments in this study) P1s choosing the most popular punishment investment versus a proportion of 0.1 P1s choosing another punishment investment. This suggests that we have insufficient punishers for this analysis in all treatments in the inefficient punishment condition (see Table 5.2).

5.7.5 Supplementary analysis (i)

When punishment was efficient, in both treatments where P2 stealing did not create disadvantageous inequality for P1 (treatments A & B), if P1 used punishment, the modal punishment investment chosen was $0.20. This punishment investment inflicted the largest costs on P2 and thus maximized advantageous inequality for P1. In treatment A, significantly more P1s invested $0.20 into punishing a stealing P2 than invested $0.15 or $0.10, though the difference between the proportion investing $0.20 versus $0.10 did not remain significant following p-value adjustment for multiple comparisons (Benjamini & Hochberg, 1995; Fisher's exact test; Table 5.S2). There was no significant difference in propensity to invest $0.20 versus $0.05 (Fisher's exact test, Table 5.S2). In treatment B, significantly more P1s with efficient punishment invested $0.20 into punishing a stealing P2 than invested $0.10 or $0.15. In treatment B, the difference between the number of P1's with efficient punishment that invested $0.20 versus $0.05 was non-significant (Fisher's exact test, Table 5.S2).

Treatment	Proportion choosing maximum punishment investment ($0.20) ± SE	Comparison (proportion ± SE)	P-value	P^{BH}-value	n
A	0.46 ± 0.1	$0.05 (0.32 ± 0.09)	0.412	0.495	28
		$0.10 (0.18 ± 0.07)	0.044	0.066	
		$0.15 (0.04 ± 0.04)	0.001	0.002	
B	0.50 ± 0.11	$0.05 (0.35 ± 0.11)	0.523	0.523	20
		$0.10 (0.05 ± 0.05)	0.003	0.001	
		$0.15 (0.10 ± 0.07)	0.014	0.028	

Table 5.S2 The proportion of players who chose the maximum punishment investment compared to the other punishment investments in treatments A & B. Data are restricted to players that punished a stealing partner and had access to efficient punishment. P values are generated from two-sided Fisher's exact tests. Benjamini-Hochberg adjusted P^{BH}-values are also presented to account for multiple comparisons. The final column shows the total sample size (n) for the comparisons.

5.7.6 Supplementary analysis (ii)

Although both the current study and R&M used a 1:3 fee to fine, the minimum cost for punishment in the current study was $0.05 whereas the minimum cost in R&M was $0.10. Thus, we wanted to test whether the increased cost to punish a stealing partner in R&M might have explained the decreased tendency to punish, particularly when P2 stole but this did not result in DI. To test whether the contradicting findings of Raihani & McAuliffe (2012b) and this study could be explained by the cost of punishment, we compared the proportion of P1 in the efficient punishment condition that chose a non-zero punishment investment when (i) P2 didn't steal; (ii) P2 stole but the stealing did not result in disadvantageous inequality for P1 ('P2 stole no DI'); (iii) and P2 stole resulting in disadvantageous inequality for P1 ('P2 stole DI'), restricting data to P1's who spent $0.10 or more on punishment.

P1 was more likely to invest $0.10 or more on punishing when P2 stole compared to when P2 didn't steal; even when stealing did not result in disadvantageous inequality (Table 5.S3; Figure 5.1). Moreover, the tendency to invest $0.10 or more in punishment was increased when stealing resulted in disadvantageous inequality (proportion punishing (i) non-stealing P2 = 0.02 ± 0.01; (ii) stealing P2, no DI = 0.14 ± 0.02; stealing P2, DI = 0.28 ± 0.02; Table 5.S3; Figure 5.1). These results suggest that the difference in the findings of Raihani & McAuliffe (2012b) and this study cannot be explained by the cost of punishment.

Punishment condition	**Comparison**	**P-value**	**P^{BH}-value**	**n**
Efficient	P2 didn't steal vs. P2 stole no DI	<0.001	<0.001	831
	P2 didn't steal vs. P2 stole DI	<0.001	<0.001	944
	P2 stole no DI vs. P2 stole DI	<0.001	<0.001	577

Table 5.S3 The P-values (and Benjamini-Hochberg adjusted P^{BH}-values) generated by Fisher's exact tests (two-sided) comparing the proportion of P1 in the efficient punishment condition that chose a $0.10 or more punishment investment across the different conditions.

Chapter 6

Exploring the Motivations for Punishment: Framing and Cross-Cultural Effects

6.1 Note

This work is currently in preparation for submission. Nichola Raihani contributed to experimental design and discussion. Katie McAuliffe contributed to discussion. I designed the experiment, collected the data, analysed the data and wrote the paper.

6.2 Abstract

Is punishment motivated by the desire to reciprocate losses ('revenge') or by the desire to reduce payoff asymmetries between the punisher and the target? In many experiments these two possibilities are impossible to separate because punishers typically experience losses and disadvantageous inequality when they interact with a cheat. Recent work that has separated these two possible motivations has suggested that punishment is more likely to be motivated by experiencing disadvantageous inequality than by a desire for revenge. Nevertheless, these findings do not replicate across different studies. Here, we suggest that considering culture - previously overlooked as a possible source of variation in responses - is important for understanding when and why individuals punish one another. We conducted a two-player stealing game with punishment, using subjects recruited from the USA and India. US-based subjects were sensitive to both losses and disadvantageous inequality when deciding whether to punish a partner. Antisocial punishment (paying to reduce the income of non-stealing partners) was vanishingly rare among the US-based subjects. India-based subjects, on the other hand, punished at higher levels than US-based subjects and, so long as they did not experience disadvantageous inequality, punished stealing and non-stealing partners indiscriminately. Nevertheless, as in the USA, when stealing resulted in disadvantageous inequality, India-based subjects punished stealing partners more than non-stealing partners. These results are consistent with previous studies that have demonstrated cross-cultural variation in punitive behaviour and support the idea that under some circumstances punishment might function to improve relative status, rather than to enforce cooperation.

6.3 Introduction

The factors underpinning decisions to cooperate and to punish others have traditionally been studied in stylized economic games in laboratories, mostly using Western undergraduates as the representative sample (Fehr & Gächter 2002; Camerer 2003; Dawes et al. 2007; Henrich et al. 2010). One traditional paradigm is the public goods game (Ledyard 1995), where group members' contributions are costly in that they benefit the group at the expense of the contributor (Fehr & Gächter 2002). In such settings, it has been shown that subjects are willing to forfeit a portion of their endowment to fine uncooperative players; a behaviour that is interpreted as punishment (Fehr & Gächter 2002). Individuals that interact with cheating partners often experience negative emotions, and the strength of these emotions has been linked to the propensity to invest in costly punishment (Fehr & Gächter 2002; Sanfey et al. 2003; Xiao & Houser 2005; Grimm & Mengel 2011; Wang et al. 2011). Nevertheless, these negative emotions could have one of two (not mutually exclusive) sources (Raihani & McAuliffe 2012a). First, since cooperators that interact with cheats incur losses relative to those that interact with cooperators, the desire to punish might reflect the disutility associated with experiencing lower than expected payoffs and an associated desire to inflict reciprocal harm on a cheating partner (also termed the desire for 'revenge'; McCullough et al. 2013). It is argued that the functional significance of punishment motivated by revenge is to deter cheating interaction partners from repeating their harmful actions again in subsequent interactions with the punisher (McCullough et al. 2013). Evidence from humans and non-human species suggests that punishment might often achieve this deterrent function (Fehr & Gächter 2002; Raihani et al. 2010; Raihani et al. 2012 but see Dreber et al. 2008; dos Santos et al. 2013). However, in stereotypical laboratory games, cheats typically end up with higher payoffs than cooperators meaning that, in addition to losses, individuals also experience disadvantageous inequality when interacting with a cheating partner. This means that the desire to punish could stem from disadvantageous inequality aversion (the disutility associated with experiencing lower payoffs than others, (Fehr & Schmidt 1999). In many of the most common games used to explore punishment in humans, these two possible motives are

impossible to separate. It has been suggested punitive sentiments that are motivated by disadvantageous inequality aversion (rather than desire for revenge) are more consistent with the idea of a fitness-levelling (rather than a deterrent) function of punishment (Price et al. 2002).

In an effort to identify the motive underpinning punishment decisions, recent studies have placed subjects in conditions where they experience either inequality or losses independent of one another (e.g. Dawes et al. 2007; Raihani & McAuliffe 2012b; Bone & Raihani 2015). Based on these approaches, empirical data have supported the idea that punishment might be motivated more by inequality aversion than by the desire for revenge. For example, in a random income game (where no cheating occurred and the desire to inflict reciprocal harm could therefore be ruled out), Dawes et al. (2007) showed that subjects would pay a cost to reduce the income of the richest players in the group (and also to increase the earnings of the poorer members). A follow-on study showed that the preference for equal outcomes in the random income game was significantly associated with willingness to punish cheats in a public goods game, leading the authors to conclude that egalitarian motives underpin the desire to punish cheats in social interactions (Johnson et al. 2009). Other supportive evidence includes the finding that third-party punishment decisions, which - by definition - cannot be motivated by a desire for revenge, are positively linked to the degree of envy felt by the third-party with respect to the transgressor's gains (Pedersen et al. 2013; but see Jordan et al. 2014).

In a recent study, Raihani & McAuliffe (2012b) provided a more direct empirical test of the relative importance of losses versus inequality in predicting punishment. The test was based on a two-player game, where player one was initially endowed with $0.70 and player two's endowment varied according to one of three treatments (A: $0.10; B: $0.30; C: $0.70). In the first stage, player 2 (P2) was given the option to steal $0.20 of Player 1's (P1) endowment. In the second stage of the game, P1 could pay a fee ($0.10) to punish P2 (by $0.30). In this setup, P1 therefore experienced the same losses if P2 decided to steal ($0.20 in each treatment) but different relative outcomes (A: advantageous inequality; B:

equal outcomes; C: disadvantageous inequality). Thus, the design allowed the possibly separate effects of incurring losses versus experiencing disadvantageous inequality on P1's decision to punish P2 to be disentangled. When P2 stole but stealing did not result in disadvantageous inequality, about 15% of P1 punished P2, which was not significantly different from the number that punished when P2 didn't steal. However, when P2 stealing created disadvantageous inequality for P1, more than 40 % of P1 punished when P2 stole; a significantly higher proportion than when P2 didn't steal. Thus, the study supported the idea that punishment decisions are motivated primarily by disadvantageous inequality rather than by a desire for revenge *per se*.

In an attempt to replicate and extend the earlier Raihani & McAuliffe (2012b) study (using a similar game structure), Bone & Raihani (2015) (Chapter 5 of this thesis) showed that players were more likely to punish stealing partners if stealing resulted in disadvantageous inequality rather than equality or advantageous inequality (Bone & Raihani 2015). However, they also found that P1s punished stealing P2s more than non-stealing P2s, even when stealing did not result in disadvantageous inequality. The findings from the more recent study suggest that punishment stems from both a desire for revenge and an aversion to inequality and therefore seem to contradict the results from the earlier study. We designed the current study to tease apart possible explanations for the differences between the findings of the original Raihani & McAuliffe (2012b) study (hereafter R&M) and the more recent Bone & Raihani (2015) study (hereafter B&R); and to get a better understanding of whether motives based on revenge or inequality aversion, respectively, are a more important driver of the decision to punish.

Our initial hypothesis was that methodological details could explain the discrepancy in the results of the two studies. In R&M, subjects were initially endowed with $0.70 which, if they were paired with a stealing partner, was reduced to $0.50. In contrast, in B&R, subjects started out with $1.10, which would be reduced to $0.90 if the partner stole $0.20. Although the losses experienced were consistent across the two studies ($0.20 in both), it is possible that the loss was perceived as greater in B&R due to a phenomenon called the

left-digit effect (Thomas & Morwitz 2005), where losses that result in the left digit changing (i.e. from $1.10 to $0.90) are perceived as larger than those where the left digit does not change (i.e. as in $0.70 to $0.50). A second possibility is that variation in the cost of punishing a stealing partner (relative to the punisher's endowment) might have been responsible for the different findings. Although the same fee to fine ratio (1:3) was used in both studies, punishment in R&M cost twice as much as in B&R ($0.10 versus $0.05) meaning that, in B&R, punishers both had a higher endowment to spend on punishment and could punish a partner more cheaply. Variation in the cost of punishing has previously been shown to influence whether people use this option when it is available (Anderson & Putterman 2006). The current study was therefore designed to try and identify whether either of these methodological details might have systematically affected our results.

We also wished to explore the possibility that cultural differences between players might have affected punishment strategies. Data for both previous studies were collected via the online crowdsourcing platform, Amazon Mechanical Turk (MTurk; www.mturk.com), where the vast majority of workers hail from either the USA or India (Ross et al. 2010). While R&M recruited participants from both countries, B&R restricted participation to subjects based in the USA. Importantly, R&M did not control for possible cross-cultural differences in subjects' behaviour. If there are systematic differences in the way that India-based versus US-based subjects behave in economic games, particularly with respect to punishment, then the different demographic sampling could explain the different results we saw across the two studies. Previous work has shown stark cross-cultural differences in the propensity to punish, both when punishment is aimed at social cheats (Henrich et al. 2006; Marlowe & Berbesque 2008) and when punishment is aimed at non-cheating or overtly cooperative individuals (commonly referred to as 'antisocial punishment'; Ellingsen et al. 2012; Herrmann et al. 2008; Gächter & Herrmann 2009). If we were to find differences in propensity to punish under different conditions between the US-based and India-based players on MTurk then this might help us to understand the discrepancies in our results of and, more generally, to gain insights into the factors that motivate individuals to punish.

6.4 Methods

6.4.1 Experimental protocol

This project was approved by the University College London ethics board under the project number 3720/001. Prior to taking part in the study, subjects were required to tick a box to indicate that they understood that they were taking part in scientific research and that their participation was voluntary. No deception was used in this study and participants were not debriefed as to the purpose of the study after the game. All data were collected in November 2014 using Amazon Mechanical Turk (see General Methods for details and justification for online data collection using MTurk). Using MTurk allowed us to recruit a more diverse demographic sample than the typical western, educated, industrialized, rich and democratic (WEIRD; Henrich et al. 2010) samples used in the majority of behavioural experiments (Buhrmester et al. 2011). Subjects were recruited from 24 countries (see supplementary materials), with the vast majority (95 %) hailing from USA (n = 1941) or India (n = 315). We recruited workers to play a modified version of the game used in R&M and B&R (described above and see Supporting Information S1 for instructions given to players). MTurk workers are identified by a unique 14-digit worker ID rather than their names (Mason & Suri 2012). Workers were told that their ID would not be revealed to their partner in the game, thus ensuring anonymity. Of the 2,392 workers recruited to play the game, 1,196 were randomly assigned to the role of 'player 1' (P1) and the remaining 1,196 to the role of 'player 2' (P2). To be eligible to participate in the study, all players had to answer correctly three comprehension questions about the game. Prior to taking part in the game, subjects were first asked to provide some background demographic information on their age, gender, education, income levels and country of origin (see Supplementary materials; Table 6.S1).

In the game, P1 was allocated a bonus of either $0.70, $1.10 or $1.30, depending on the experimental treatment (Table 6.1). P2 was also allocated a bonus according to the scenarios outlined in Table 6.1. Thus, we had three different experimental treatments (corresponding to different starting amounts for P1) and within each treatment players were allocated to one of four scenarios

(corresponding to the different starting amounts for P2). The game consisted of two stages. In the first stage, players were first informed of their own initial endowment and the endowment of the partner; and P2 was then given the choice to take $0.20 of P1's bonus or to do nothing (as in R&M). In the second stage, P1 was given the option to punish P2 (framed as 'reducing P2's bonus' in the game instructions; see Supplementary materials), by paying $0.10 to reduce P2's bonus by $0.30 (as in R&M). Players were matched with partners ex-post (as in (Rand 2012; Raihani & McAuliffe 2012b; Bone & Raihani 2015). The four scenarios for each treatment were chosen to allow us to identify different possible motives underpinning P1's decision to punish P2. In all scenarios, P1 incurred a loss of $0.20 if P2 stole (leading to a second stage bonus of $0.50, $0.90 or $1.10 in Treatments A, B and C, respectively, Table 6.1). For each treatment, in scenarios (i - iii) P1 started out better off than P2, but in scenario (iv) P1 and P2 started with equal initial bonus. In each scenario if P2 decided to steal from P1, the following outcomes ensued:

scenario (i): P1 remained better off than P2;

scenario (ii): P1 and P2's final outcomes were equal;

scenarios (iii): P1 ended up worse off than P2;

scenario (iv): P1 ended up worse off than P2.

Treatment	P1's endowment	Scenario	P2's endowment	Outcome if P2 stole (P1-P2)	Sample size (P2 stole)	Sample size (P2 did not steal)
A	$0.70	i	$0.10	$0.50 - $0.30	49	50
		ii	$0.30	$0.50 - $0.50	50	50
		iii	$0.50	$0.50 - $0.70	50	50
		iv	$0.70	$0.50 - $0.90	50	50
B	$1.10	i	$0.50	$0.90 - $0.70	49	50
		ii	$0.70	$0.90 - $0.90	50	50
		iii	$0.90	$0.90 - $1.10	49	50
		iv	$0.90	$0.90 - $1.30	50	50
C	$1.30	i	$0.70	$1.10 - $0.90	50	50
		ii	$0.90	$1.10 - $1.10	50	50
		iii	$1.10	$1.10 - $1.30	50	50
		iv	$1.30	$1.10 - $1.50	50	49

Table 6.1 Initial endowments allocated to P1 and P2, the outcome if P2 stole and the sample size of P1's who interacted with a stealing / non-stealing P2 according to the treatment and scenario.

Thus, the four scenarios for each treatment allowed us to disentangle the possibly separate effects of experiencing losses versus experiencing disadvantageous inequality on P1's decision to punish P2. We included both scenarios (iii) and (iv) to further delineate two potential motivations that might have driven P1 to punish a stealing P2. First, punishment might be motivated by frustration that the starting bonuses had initially been equal but were now tipped in favour of P2 (scenario iv). Alternatively, the decision to punish might be insensitive to the initial starting bonuses and instead just reflect P1's frustration at a disadvantageous outcome (scenario iii).

While the four scenarios allowed us to explore how different outcomes affected P1's decision to punish P2, the three treatments (A, B and C) allowed us to control for the possibility that the initial starting bonus of P1 might affect punishment decisions. Specifically, Treatment A (where P1's starting bonus was $0.70) was a direct replication of R&M. Treatment B (P1's starting bonus $1.10) was included to test whether the left-digit effect (Thomas & Morwitz 2005) might have motivated P1 to punish a stealing P2 (i.e. if losses were perceived as being greater when moving from $1.10 to $0.90 than from $0.70 to $0.50). Treatment C (P1's starting bonus $1.30) was included to investigate the possibility that P1 would be more likely to invest in punishment when the cost of punishment was smaller, relative to the initial endowment. Thus, if either the left-digit effect or the cost of punishment (as a proportion of the endowment) were affecting P1's punitive behaviour, then we would have expected the starting bonus to have had an effect on P1's decision to punish P2.

6.4.2 Analysis

First, we used our data to ask whether the left-digit effect or the relative cost of punishment might have affected P1's decision to punish a stealing P2. If the left-digit effect was increasing P1's tendency to punish P2, then we should have detected increased punishment of a stealing P2 in treatment B (starting bonus $1.10) relative to treatments A or C, even when stealing did not result in disadvantageous inequality for P1. Conversely, if P1 was more likely to punish a stealing partner when punishment was a smaller cost, relative to the initial

endowment, then we should have seen increased punishment in treatment C, relative to A or B. We explored these relationships using a chi-squared test. Data were restricted to instances where P2 stole from P1 and when stealing did not result in disadvantageous inequality for P1 (n = 297).

Next, we explored how country of origin and outcome inequality affected P1's punishment decisions. We used a generalised linear model (GLM), with the term 'punish' set as a binary response term (1 = P1 punished P2 = 1; 0 = P1 did not punish P2). As in R&M and B&R, we were interested in whether experiencing losses or disadvantageous inequality had a greater effect on P1's decision to punish P2. To this end, we used our data to create an explanatory term called 'outcome' that captured this variable. The term 'outcome' was a 3-level categorical variable with the levels ‘P2 didn’t steal' (P2 did not steal from P1); 'P2 stole no DI' (P2 stole but the stealing did not result in disadvantageous inequality); and 'P2 stole DI' (P2 stole and the outcome was disadvantageous inequality for P1). We included the following explanatory terms in the GLM: 'outcome'; 'equality ruined' (1 = P2 stealing resulted in a formerly equal outcome becoming unequal; 0 = otherwise); and, 'country' (a 2-level factor specifying P1's country of origin: India / USA). We also included the following two-way interaction: 'country x outcome'. Because our preliminary chi-squared analysis (described above) showed that the starting bonus was not an important predictor of P1's punishment decision, we combined data for all treatments (A, B and C) for this analysis. Since we were interested in exploring the possibility that culture could affect subjects' decisions in the game, and the vast majority (95 %) of our subjects allocated to the P1 role were recruited from either the USA (n = 962) or from India (n = 176), we restricted data to responses from these subjects. The remaining 58 subjects allocated to the P1 role hailed from 24 different countries (see Supplementary materials; Table 6.S1), meaning that we did not have a sufficient sample size to make meaningful inference about country-level effects on behaviour for these individuals. We therefore had a sample size of 1138 P1 punishment decisions for this model. All proportions are reported with 95 % confidence intervals. All data were analysed using R version 3.0.3 (www.r-project.org). For the GLM we used an information-theoretic approach with model averaging, as described in Grueber

et al. (2011), to determine the relative importance of the explanatory terms included in each model. The input variables were centred by subtracting the mean (Schielzeth 2010), this allows averaging over models that include different interaction terms (see General Methods for details).

6.5 Results

Neither the left-digit effect nor the cost of punishing relative to the initial endowment seemed to affect P1's punishment decisions, in scenarios where stealing did not result in disadvantageous inequality for P1. In treatment A (starting bonus $0.70), P1 punished a stealing partner on 12 / 99 (12.1 %) occasions; in treatment B (starting bonus $1.10), P1 punished a stealing partner on 16 / 99 (15.0 %) occasions; and in treatment C (starting bonus $1.30), P1 punished a stealing partner on 10 / 99 (10.2 %) occasions. These differences across treatments were not significant at conventional levels (Chi-squared test: X = 1.69, df = 2, P = 0.43).

On average, India-based subjects were more punitive than US-based subjects (Table 6.2). In addition, subjects from India and the US responded differently to outcome (the variable describing whether the partner stole; and whether this resulted in disadvantageous inequality if so, Table 6.2; Figure 6.1). Subjects from India were more likely than US-based subjects to punish a non-stealing partner (proportion India-based P1 punishing non-stealing P2 = 0.18 (0.11, 0.28); US-based P1 punishing non-stealing P2 = 0.01 (0.00, 0.02); Figure 6.1). So long as stealing did not result in disadvantageous inequality, India-based subjects did not punish a stealing partner more than a non-stealing partner. In other words, Indian subjects punished 'antisocially' at higher levels than US-based players and did not seem to increase punishment in response to experiencing losses in the absence of disadvantageous inequality (Figure 6.1b). Nevertheless, when stealing did result in disadvantageous inequality, India-based subjects were even more likely to punish the partner (proportion India-based subjects P1 punishing stealing P2 when stealing did not result in disadvantageous inequity = 0.19 (0.11, 0.32); when stealing did result in disadvantageous inequity = 0.50 (0.34, 0.66); Figure 6.1b). The patterns for US-based subjects were different: even when stealing did not

result in disadvantageous inequality, subjects were more likely to punish stealing than non-stealing partners (Figure 6.1a). Nevertheless, US-based subjects were also sensitive to inequality and punished stealing partners even more when stealing resulted in disadvantageous inequality (proportion US-based P1 punishing stealing P2 when stealing did not result in disadvantageous inequity = 0.11 (0.08, 0.16); when stealing did result in disadvantageous inequity = 0.27 (0.23, 0.34); Figure 6.1a). Thus, Indian subjects only punished stealing partners more than non-stealing partners when stealing resulted in disadvantageous inequality; whereas US subjects were sensitive to both losses and outcome inequality when making a punishment decision (Table 6.2).

Although the term 'equality ruined' was a component of the top models, the confidence intervals for this term spanned zero indicating 'equality ruined' was not an important driver of P1's punishment decision (Table 6.2).

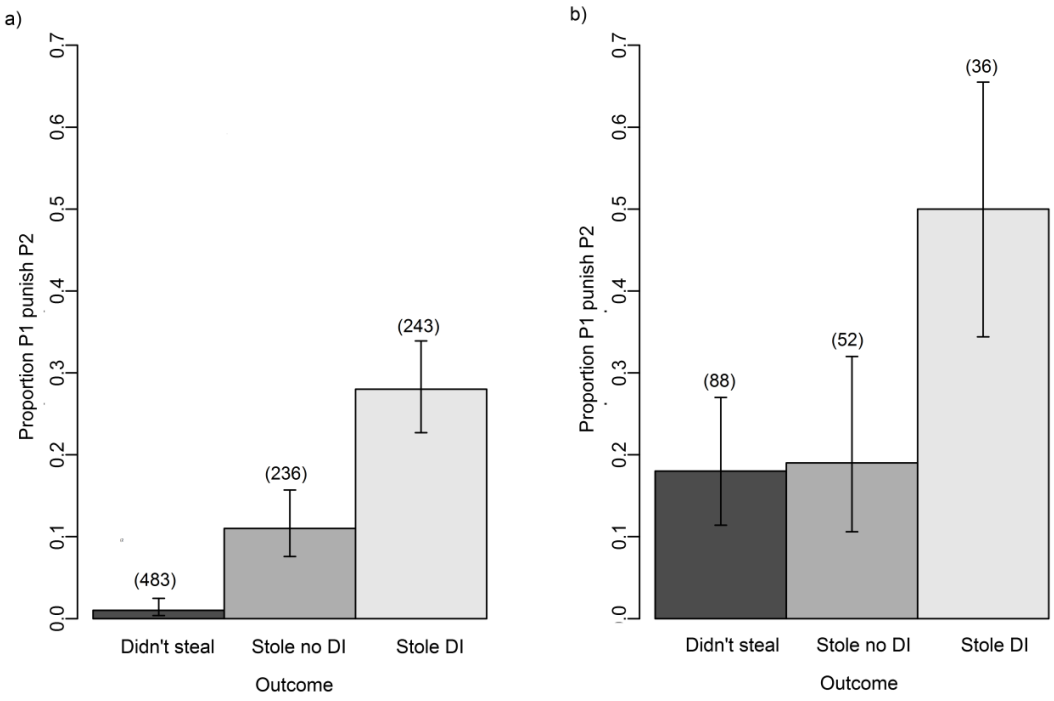


Figure 6.1 The proportion of P1 who punished P2 when P2 didn't steal ('Didn't steal'), P2 stole but the stealing did not result in disadvantageous inequality ('Stole no DI') or P2 stole and the outcome was disadvantageous inequality for P1 ('Stole DI'). Data are shown for players based in **a)** the USA and **b)** India. Error bars show the 95 % confidence intervals. Sample sizes for each condition are indicated in parentheses. Plots are generated from raw data.

Parameter	Estimate	Unconditional SE	Confidence Interval	Relative Importance
Intercept	-2.75	0.20	(-3.14, -2.36)	
Country (India / USA)	-1.93	0.30	(-2.52, -1.35)	1.00
Outcome				1.00
P2 stole no DI	2.10	0.42	(1.27, 2.93)	
P2 stole DI	3.31	0.41	(2.51, 4.12)	
Outcome x Country				1.00
P2 stole no DI	2.40	0.67	(1.09, 3.71)	
P2 stole DI	2.10	0.64	(0.85, 3.36)	
Equality ruined	-0.05	0.17	(-0.69, 0.34)	0.31

Table 6.2 Estimates, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models. All input variables were centred by subtracting the mean (Schielzeth 2010). Standard errors are unconditional, meaning that they incorporate model selection uncertainty. Outcome is a 3-level categorical variable: 'P2 didn't steal' = player 2 did not steal; 'P2 stole no DI' = player 2 stole but this did not result in disadvantageous inequality for P1; and 'P2 stole DI' = player 2 stole and this resulted in in disadvantageous inequality for P1. For outcome, 'P2 didn't steal' was the reference level. Estimates from the same model when the reference level for outcome is ' P2 stole no DI' presented in Supplementary Information.

6.6 Discussion

This study was designed to try and explain why we found seemingly contradictory evidence for the importance of revenge as a motive underpinning punishment decisions across two separate studies. The R&M study found that only experiencing disadvantageous inequality (rather than losses in the absence of inequality) motivated subjects to punish a stealing partner. In the more recent B&R study, we found that both losses and inequality motivated punishment decisions. Based on the current dataset, it seems unlikely that minor methodological differences (specifically, differences in stake size or the relative cost of punishment) across the two studies can explain this discrepancy in the respective studies. Instead, we suggest that differences in the tendency to punish a stealing partner are largely explained by subjects' country of origin. For US-based subjects, the decision to punish was affected both by whether the partner stole and by whether the punisher experienced outcome inequality. Thus, the motives underpinning punishment apparently stemmed both from a desire for revenge and from the disutility associated with experiencing relatively lower payoffs than a partner (a finding replicated by Bone & Raihani 2015). US-based subjects were very unlikely to punish a non-stealing partner. This pattern was not replicated for the Indian subjects; punishment aimed at non-stealing individuals was an order of magnitude higher. Where stealing did not result in disadvantageous inequality, there was no effect of losses on the propensity that Indian subjects would punish. By contrast, punishment was increased when stealing resulted in disadvantageous inequality for the punisher. The findings from a reanalysis of the data from the R&M study (separating US and India-based subjects) were consistent with these overall patterns (see supplementary materials for details).

Our data showed that a non-negligible proportion of players were prepared to incur a cost to harm the partner, even when the partner did not steal any money (Figure 6.1). The tendency to pay to harm a non-stealing (or even an overtly cooperative) partner has been observed in several other studies (e.g. Herrmann et al. 2008; Gächter & Herrmann 2009; Anderson & Putterman 2006; Sylwester et al. 2013; Abbink & Sadrieh 2009), where it has variously been described with the

labels 'antisocial punishment' (Herrmann et al. 2008; Gächter & Herrmann 2009; Anderson & Putterman 2006; Sylwester et al. 2013; Abbink & Sadrieh 2009; Bryson et al. 2014) or 'spite' (Abbink & Sadrieh 2009). While the use of the term 'antisocial punishment' does not fit the functional definition of punishment (as a harmful action aimed at changing the target's, or a bystander's, behaviour in future interactions, (Clutton-Brock & Parker 1995; Raihani, Thornton, et al. 2012) in our one-shot game, we retain its use here to informally describe the behaviour of individuals who paid a cost to harm a non-stealing partner. Antisocial punishment was more common among the India-based than the US-based subjects, a finding which mirrors previous studies that have shown cross-cultural variation in antisocial punishment (Herrmann et al. 2008; Gächter & Herrmann 2009). Cross-cultural variation in antisocial punishment is thought to be predicted by weak rule of law, which is itself negatively linked to Gross Domestic Product (GDP; Herrmann et al. 2008; Sylwester et al. 2013; Prediger et al. 2013). According to the World Bank (2013), the GDP of the USA in 2013 was 16,800,000 million dollars, compared with 1,876,797 million dollars for India in the same year, which supports the idea that the GDP of a country will be negatively associated with citizens' propensity to punish antisocially. In addition, Herrmann et al. (2008) showed that antisocial punishment was associated with low scores for civic norms of cooperation. Following Herrmann et al. (2008) we used the questions from the most recent World Values Survey (2010-2014) to calculate mean scores for civic norms of cooperation for Indian and US citizens (see Supplementary materials). We calculated a mean score of 8.73 for US citizens (comparable to the score of 8.65 as calculated by Herrmann et al. 2008, using data from 1999-2004). For Indian citizens, we calculated a mean civic norms score of 6.84, which is lower than for any of the countries that were included in the original Herrmann et al. (2008) analysis. Thus, our results do seem to be consistent with the idea that low civic norms of cooperation and rule of law are associated with increased tendency for antisocial punishment.

In stereotypical public goods games, punishment is most often directed from cooperative individuals towards uncooperative targets. When targets are aware of who punished them and are given the chance to counter-punish, many individuals

do choose this option (Dreber et al. 2008; Nikiforakis 2008; Janssen & Bushman 2008; Fehl et al. 2012; Denant-Boemont et al. 2007; Nikiforakis & Engelmann 2008). On this basis, it has been proposed that antisocial punishment might sometimes reflect retaliation by free-riders who were punished by cooperative partners, or a pre-emptive strike against expected punishment in subsequent rounds (Cinyabuguma et al. 2006; Nikiforakis & Normann 2008; Herrmann et al. 2008; Sylwester et al. 2013; Raihani & Bshary 2015). In our study, however, retaliation (pre-emptive or otherwise) can be ruled out as a motive underpinning antisocial punishment since the punishers did not receive any previous punishment to retaliate against. We can also rule out the possibility that individuals used antisocial punishment to deter partners from punishing in subsequent rounds, since this was a one-shot interaction. Instead, our findings lend more support to the recent idea that spiteful actions (including antisocial punishment) might simply reflect aggressive competition for status (as proposed by Charness et al. 2010; Prediger et al. 2014; Sylwester & Roberts 2010; Raihani & Bshary 2015). This idea has been supported by previous studies that have shown that the tendency to punish antisocially tends to disappear when fee to fine ratios are adjusted such that punishers cannot improve their standing relative to that of the target (Falk et al. 2005 but see Egas & Riedl 2008).

It has been suggested that conceiving of punishment as an aggressive act, designed to improve relative status, can go some way to explaining cross-cultural differences in antisocial punishment use since the benefits of acquiring higher status than others might vary with socio-ecological factors such as rule of law and resource availability (for which GDP might provide a reasonable proxy; Sylwester et al. 2013; Prediger et al. 2014; Bryson et al. 2014). For example, in a recent empirical study, Prediger et al. (2014) used the 'joy of destruction' game (Abbink & Sadrieh 2009) to show that willingness to engage in costly spiteful behaviour correlates positively with resource scarcity. This empirical result supports earlier theoretical work which has shown that costly spite (aimed at competitors) can be favoured under resource (Gardner & West 2004b; Lehmann et al. 2009). If 'spiteful' or antisocial punishment functions to improve status (rather than to deter partners from cheating) it need not be motivated by a desire for revenge and might

instead be aimed rather indiscriminately against stealing and non-stealing partners. This hypothesis would, nevertheless, predict increased punishment of stealing partners (relative to non-stealing partners) when stealing results in disadvantageous inequality. The empirical data from our India-based subjects matches these predictions. Thus, our data support the idea that punishment might be used at least in some contexts to establish or maintain dominance over peers. An obvious future extension to this study would be to investigate whether we would observe similar punishment patterns when the fee to fine ratio is 1:1 (or lower) such that punishers cannot improve their payoffs relative to those of targets. If the hypothesis that punishment sometimes represents aggressive competition for status is correct, then we would expect a marked decrease in antisocial punishment when punishers cannot improve their payoff relative to that of the target. It would also be interesting in any follow up to obtain self-reports from subjects on the emotions they feel prior to making their punishment decision. While some studies have suggested that punishment might often be preceded by negative emotions, such as anger or disgust (e.g. Fehr & Gächter 2002; Sanfey et al. 2003), this might not necessarily be the case when punishment is motivated by a competitive drive to elevate relative status and is used indiscriminately against prosocial and antisocial partners.

While the data from the India-based subjects provides some support for the idea that punishment might be proximately driven by competitive motives, the data from the US-based subjects suggest that revenge based motives cannot easily be ruled out. A similar conclusion was also reached based on the data collected for the Bone & Raihani (2015) study, where it was shown that individuals would invest in 'inefficient' punishment (fee to fine: 1:1) if this was the only option available but that, when given access to an 'efficient' punishment option (fee to fine: 1:3), typically invested the amount that created equal outcomes for the punisher and the target. Investment in inefficient punishment supports the idea that punishment is motivated by a desire for revenge rather than by competitive motives to equalise or increase payoffs relative to the target (since, by definition, inefficient punishment cannot have any bearing on relative payoffs of punisher and target). Nevertheless, the preference to equalise outcomes when this was

possible supports the idea that punishment might be motivated by egalitarian preferences and therefore be more likely to serve a fitness-levelling function. If the assumption that motives underpinning decisions can lend some insight into the likely evolved function of the behaviour is correct, then - based on the previous data and the data collected for the current study - we suggest that punishment might in fact serve both a deterrent function (motivated primarily by revenge) and a fitness-levelling (or improving) function (motivated by competitive desire; which may be increased when punisher experiences disadvantageous inequality). The relative importance of the two functions (and associated motives) might be expected to vary according to context (e.g. Gardner & West 2004). Based on the current data, it appears that culture is likely to play an important role in determining the relative importance of the two functions (and underlying motives) of punishment, although it is not yet clear which of the many factors that vary across cultures might be causal. Moreover, whether among-culture variation in punishment patterns is greater than that which is observed within cultures remains an open avenue for investigation.

To conclude, the discrepancies that we observed across two similar studies concerning the importance of revenge as a proximate motive underpinning punishment decisions do not seem to reflect minor methodological details or, as was even more of a concern, the general unreliability or noisiness of data collected via the online crowdsourcing platform, Mechanical Turk. Instead, we suggest that the differences between the studies reflect different demographic sampling and the fact that cultural differences in punitive behaviour were not accounted for in the earlier R&M study. In fact, comparing the data from our US-based players in this study with the comparable data obtained by Bone & Raihani (2015) reveals remarkable consistency in the overall patterns (and, in fact, others have also reported similar levels of consistency across studies conducted using MTurk, Peysakhovich et al. 2014). There is much more to be done in understanding what motivates punishment decisions and, even more importantly, why these motivations vary across contexts.

6.7 Supplementary materials

6.7.1 Experimental instructions

Having completed the demographic questions, participants were redirected to an external survey website (https://opinio.ucl.ac.uk) to take part in the experiment. Below is a text transcription of the game instructions received by players, including comprehension questions. The example given is for a player 1 in treatment A, scenario (i); however, the general procedure is similar for all treatments and punishment conditions

Screen 1: You are about to take part in an academic study which is run by Raihani Lab based at University College London. By continuing with the HIT you are consenting to allow Raihani Lab to use your responses in the study for academic purposes. All data are anonymous (your name or worker ID will not appear in any publication related to this study). Please tick 'I agree' if you agree to these conditions. If you do not wish to participate, or if you change your mind during the course of the study, please return to the Mechanical Turk Interface and click 'Return HIT'

Screen 2: Please enter your Worker ID. Your Worker ID is needed to ensure you get your bonus. Please DO NOT enter your email address or name in this box. If you don't know your Worker ID you can find it out by opening the following page in a new window: https://www.mturk.com/mturk/dashboard

Screen 3: You are playing a game with another worker. Your worker ID will not be revealed to the other player and you will not find out their worker ID. You are Player 1. You have been allocated $0.70 bonus. Player 2 has been allocated $0.10 bonus. This game has 2 stages. Stage 1: Player 2 can choose to take $0.20 from your bonus or to do nothing. Stage 2: You find out what Player 2 did in Stage 1 and can then choose one of the following options:

- [A] reduce Player 2's bonus by paying $0.10 to reduce Player 2's bonus by $0.30. - [B] do nothing (no cost to you or Player 2)

Screen 4: Before proceeding, please answer the following questions. Please answer carefully - you will not be able to proceed if you get an answer wrong. Reminder: You are Player 1. You have a $0.70 bonus. Player 2 has a $0.10 bonus.

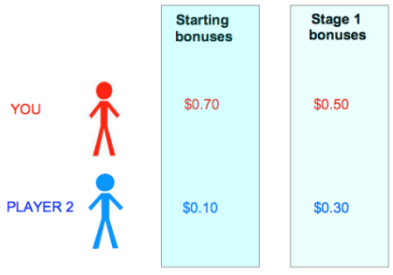
1. If Player 2 takes $0.20 from your bonus in Stage 1 how much will you have at the start of Stage 2? (I will have $0.50 / I will have $0.70 / I will have $0.90)

2. If Player 2 takes $0.20 from your bonus in Stage 1, how much will Player 2 have at the start of Stage 2? (Player 2 will have $0.70 / Player 2 will have $0.30 / Player 2 will have $0.90)

3. How much do you have to pay to reduce Player 2's bonus in Stage 2? (Nothing - it is free / it costs me $0.50 to reduce Player 2's bonus / It costs me $0.10 to reduce Player 2's bonus)

Screen 5: Well done - you got all the questions right. Click 'continue' below to continue to the game. (Continue)

Screen 6: STAGE 1. Player 2 could choose to take $0.20 of your bonus or to do nothing. Player 2 took $0.20 of your bonus. The starting bonuses and current bonuses are shown below.



Screen 7: STAGE 2. You may now pay $0.10 to reduce Player 2's bonus by $0.30. The possible final bonuses for you and Player 2 are shown below. Do you want to reduce Player 2's bonus? (Yes - I want to reduce Player 2's bonus / No - I don't want to reduce Player 2's bonus)

Screen 8: That's the end of the game. The mystery word is 'FISH'. Please return to the HIT and enter the word 'FISH' in the box before submitting your HIT. Thanks for playing!

6.7.2 Matching of subjects

As described in the General Methods section, it was not always possible to match players with a unique partner in the same scenario/treatment as them self. In this experiment 86.79 % of players (n = 2076) were matched with a unique partner in the same scenario and treatment as them self, 5.85 % of players (n = 140) were matched with a unique partner in the same scenario but a different treatment and 2.51 % of players (n = 60) were matched with a unique partner in neither the same scenario nor the same treatment as them self. This left 58 P2s who did not steal and 58 P1s who were told that their partner did steal. These players (4.85 % of players) could not be uniquely matched and so were matched with a partner in the same scenario and treatment as them self but who had already been matched with another player (Figure 6.S1).

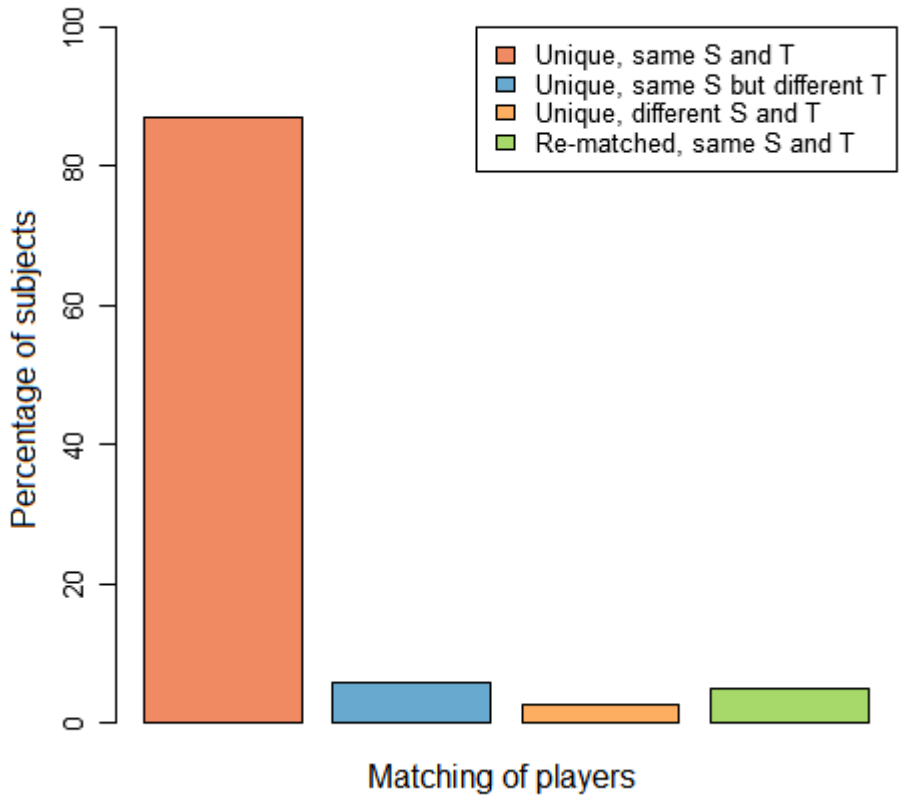


Figure 6.S1 The percentage of subjects that were matched with a unique partner in the same scenario and treatment (Unique, same S and T; Red bar), a unique partner in the same scenario but a different treatment (Unique, same S but different T; Blue bar), a unique partner in a different scenario and a different treatment (Unique, different S and T; Orange bar) or a partner in the same scenario and treatment but who has already been matched with another player (Re-matched, same S and T; Green bar).

6.7.3 Calculating norms of civic cooperation

We followed the methodology of Herrmann et al. (2008) to calculate the norms of civic cooperation for India relative to the USA from the period 2010 - 2014. We used data from the World Values Survey website (www.worldvalues.survey.org) and assessed the answers to the three questions previously analysed by Herrmann et al. 2008 to calculate the norms of civic cooperation for each country. These

three questions asked participants to answer to what extent (on a scale of 1 - 10, with 1 being never justifiable and 10 being always justifiable) the following actions could be justified: claiming government benefits to which you are not entitled (question V198); cheating on taxes if you have a chance (V201); and avoiding a fare on public transport (V199). As in Herrmann et al. (2008) we rescaled the answers such that a value of 1 would indicate weak civic norms and a value of 10 would indicate strong civic norms. In the Herrmann et al. (2008) study, the authors typically used data collected between 1999-2004; in this analysis we used the data collected from 2010-2014. Our social capital variables for norms of civic cooperation for India and the USA, respectively, were 6.84 and 8.73. Previously Herrmann et al. (2008) reported a value of 8.65 for the USA (and a global range of 6.75 - 9.81). Thus, by this metric, India appears to have extremely weak civic norms, which has been associated with increased prevalence of antisocial punishment (Herrmann et al. 2008).

6.7.4 Reanalysis of R&M data

We wanted to explore whether the cross-cultural differences in the propensity to punish found in the current study were consistent with the R&M study. In the R&M study the vast majority (81 %) of subjects allocated to the role of P1 were recruited were from the US or India. In the R&M study there was more India-based P1s (n = 130) than US-based P1s (n = 97), whereas in the current study, we had more US-based P1s (n = 962) than India-based P1s (n = 176). Although sample sizes were too small to formally test for cross-cultural differences in the R&M data, descriptive statistics support the patterns found in the current study. Subjects from India were more likely than US-based subjects to punish a non-stealing partner (proportion India-based P1 punishing non-stealing P2 = 0.16 (0.08, 0.28); US-based P1 punishing non-stealing P2 = 0.03 (0.00, 0.12); Figure 6.S1). So long as stealing did not result in disadvantageous inequality, India-based subjects did not punish a stealing partner more than a non-stealing partner. However, when stealing did result in disadvantageous inequality, India-based subjects were more likely to punish their partner (proportion India-based subjects P1 punishing stealing P2 when stealing did not result in disadvantageous inequity

= 0.20 (0.11, 0.33); when stealing did result in disadvantageous inequity = 0.36 (0.20, 0.57); Figure 6.S1b). The patterns for US-based subjects were different: even when stealing did not result in disadvantageous inequality, subjects were more likely to punish stealing than non-stealing partners (Figure 6.S1a). Nevertheless, US-based subjects were also sensitive to inequality and punished stealing partners even more when stealing resulted in disadvantageous inequality (proportion US-based P1 punishing stealing P2 when stealing did not result in disadvantageous inequity = 0.1 (0.03, 0.26); when stealing did result in disadvantageous inequity = 0.67 (0.35, 0.88); Figure 6.S1a).

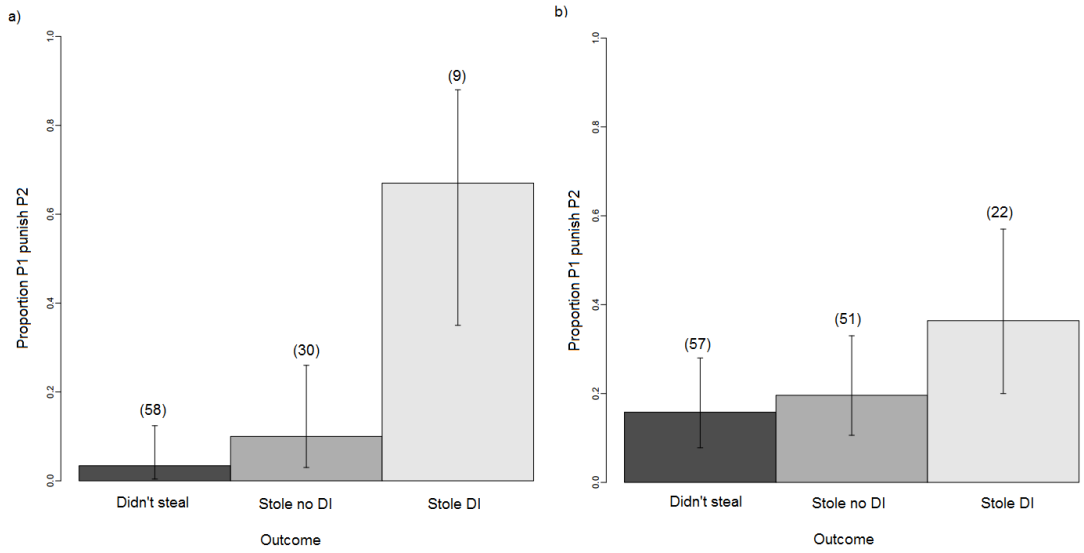


Figure 6.S2 The proportion of P1 who punished P2 when P2 didn't steal ('Didn't steal'), P2 stole but the stealing did not result in disadvantageous inequality ('Stole no DI') or P2 stole and the outcome was disadvantageous inequality for P1 ('Stole DI') in data from the R&M study. Data are shown for players based in **a)** the USA and **b)** India. Error bars show the 95 % confidence intervals. Sample sizes for each condition are indicated in parentheses. Plots are generated from raw data.

6.7.5 Supplementary tables

Parameter	**P1 demographic information** (n = 1195)
Age	Mean = 31 ± 0.3 Range = 18 - 88
Gender (n)	Females = 493 (42.0 %) Males = 680 (58.0 %) Undisclosed = 22
Country	Australia = 1 Belgium = 2 Bolivia = 1 Brazil = 1 Bulgaria = 1 Canada = 6 China = 3 Croatia = 2 Czech Republic = 1 Ethiopia = 1 Germany = 1 India = 176 Indonesia = 1 Italy = 1 Lithuania = 1 Macedonia = 1 Panama = 1 Peru = 1 Philippines = 3 Poland = 1 Romania = 2 Russia = 1 Serbia = 2 Thailand = 1 UK = 1 USA = 962

Table 6.S1 Age, gender and country of origin for all subjects allocated to role of player 1.

Parameter	**India** (n = 176)	**USA** (n = 962)
Age	Mean = 32.4 ± 0.6 Range = 19-72	Mean = 31 ± 0.3 Range = 18 - 88
Gender	Females = 57 (33 %) Males = 116 (67 %) Undisclosed = 3	Females = 420 (44 %) Males = 535 (56 %) Undisclosed = 7
Education	School = 6 (4 %) Primary degree = 112 (66 %) Graduate degree = 53 (31 %) Undisclosed = 5	School = 99 (10 %) Primary degree = 749 (79 %) Graduate degree = 100 (11 %) Undisclosed = 14
Income	Less than $12,000 = 88 (43 %) $12,000 - $24,999 = 80 (39 %) $25,000 - $49,999 = 25 (12 %) $50,000 - $99,999 = 11 (5 %) More than $100,000 = 1 (0.4 %) Undisclosed = 11	Less than $12,000 = 94 (10 %) $12,000 - $24,999 = 154 (17 %) $25,000 - $49,999 = 291 (32 %) $50,000 - $99,999 = 284 (31 %) More than $100,000 = 92 (10 %) Undisclosed = 47

Table 6.S2 Demographic information for US-based and India-based subjects allocated to role of player 1.

Parameter	Estimate	Unconditional SE	Confidence Interval	Relative Importance
Intercept	-2.75	0.20	(-3.14, -2.36)	
Country	-1.93	0.30	(-2.52, -1.35)	1.00
Outcome				1.00
P2 didn't steal	-2.10	0.42	(-2.93, -1.27)	
P2 stole DI	1.21	0.24	(0.74, 1.67)	
Outcome x Country				1.00
P2 stole no DI	-2.40	0.67	(-3.71, -1.09)	
P2 stole DI	-0.30	0.55	(-1.37, 0.77)	
Equality ruined	-0.05	0.17	(-0.69, 0.35)	0.31

Table 6.S3 Estimates, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models. All input variables were centred by subtracting the mean (Schielzeth 2010). Standard errors are unconditional, meaning that they incorporate model selection uncertainty. Outcome is a 3-level categorical variable 'P2 didn't steal' = player 2 did not steal; 'P2 stole no DI' = player 2 stole but this did not result in disadvantageous inequality for P1; and 'P2 stole DI' = player 2 stole and this resulted in in disadvantageous inequality for P1. For outcome, P2 stole no DI' is the reference level.

Chapter 7

Defectors, Not Descriptive Norm Violators, are Punished by Third-Parties

7.1 Note

This work has been published as Bone JE, Silva AE, Raihani NJ (2014) doi: 10.1098/rsbl.2014.0388. Nichola Raihani contributed to experimental design and discussion. Antonio Silva contributed to discussion. I designed the experiment, collected the data, analysed the data and wrote the paper.

7.2 Abstract

Punishment of defectors and cooperators is prevalent when their behaviour deviates from the descriptive norm. Why atypical behaviour is more likely to be punished than typical behaviour remains unclear. One possible proximate explanation is that individuals simply dislike descriptive norm violators. However, an alternative possibility exists: individuals may be more likely to punish atypical behaviour because the cost of punishment generally increases with the number of individuals that are punished. We used a public goods game with third-party punishment to test whether punishment of defectors was reduced when defecting was typical, as predicted if punishment is responsive to descriptive norm violation. The cost of punishment was fixed, regardless of the number of players punished, meaning that it was not more costly to punish typical, relative to atypical, behaviour. Under these conditions, atypical behaviour was not punished more often than typical behaviour. In fact most punishment was targeted at defectors, irrespective of whether defecting was typical or atypical. We suggest that the reduced punishment of defectors when they are common might often be explained in terms of the costs to the punisher, rather than responses to descriptive norm violators.

7.3 Introduction

Humans have a strong tendency to conform to norms of behaviour (Asch 1956; Cialdini et al. 1990; Schultz et al. 2007). In previous literature, the term ‘norm’ has been used in a rather broad way and thus the concept of norms is often split into two categories: *injunctive* norms and *descriptive* norms (Deutsch & Gerard 1955; Cialdini et al. 1990). Injunctive norms, describe beliefs about how people

ought to behave in a given situation (Irwin & Simpson 2013). Descriptive norms on the other hand, describe what behaviour is typical or is what most people do in a given situation. Conformity to descriptive norms can be an adaptive response to uncertainty regarding the appropriate behaviour in a specific context: by observing how others behave in that setting, individuals might be better able to infer what behaviour is successful (Claidière & Whiten 2012) and what is likely to be approved or disapproved by others (Cialdini et al. 1990).

Compliance with norms has been argued to underpin the existence of large-scale cooperation in human societies (Fehr & Fischbacher 2004a). Specifically, humans are thought to conform to a social norm of conditional cooperation, which is enforced by punishment of those who violate the norm (Fischbacher et al. 2001). Thus, defectors should be less likely to be punished, or be punished less severely, when they are in the majority rather than the minority. Some evidence exists to support this idea. For example, third-party punishment of defectors in a Prisoner's Dilemma Game is more severe when the partner cooperates than when both players defect (Fehr & Fischbacher 2004b). Similarly, individuals in public goods game are more likely to be punished the more their contribution deviates from the group average (Fehr & Gächter 2002; Irwin & Horne 2013).

While these findings have been interpreted as evidence that punishment is motivated by a dislike of descriptive norm deviants, we suggest an important alternative explanation: individuals are more likely to punish atypical defectors because this is by definition cheaper than punishing defectors when defection is common. In most previous studies, this explanation for the punishment of atypical behaviour has not been ruled out because the costs of punishment increase with the number of individuals that are punished (e.g. Fehr & Gächter 2002; Irwin & Horne 2013). We used a public goods game (PGG) with third-party punishment and experimentally manipulated the number of cooperators and defectors to test whether punishment is aimed specifically at descriptive norm deviants or, more generally, at defectors, when there is no additional cost to punishing the majority. We also measured the third parties desire to exclude individuals from a subsequent PGG game as an indicator of social rejection.

7.4 Methods

Data were collected in March 2014. We recruited 1050 subjects (664 males, 380 females, 6 unspecified) for our experiment using the online labour market, Amazon Mechanical Turk (AMT; www.mturk.com). Subjects were all based in the USA. We used a PGG to test whether punishment was motivated by the descriptive norm violation in this setting. Players were randomly allocated to the role of Player 1 - 4 (n = 840) or to the role of Player 5 (n = 210). Players 1 - 4 played a PGG while Player 5 was an observer who could choose to punish any or all of the four PGG players after they made their contributions. After the game, all subjects were required to fill in a questionnaire to provide demographic information (Table 7.S2).

In the PGG, Players 1 - 4 were allocated an investment token and informed that they could invest this in a 'public investment opportunity' or a 'private investment opportunity'. Public investments yielded $0.20 to the investor and $0.20 to each of the other players. Conversely, private investments yielded $0.30 to the investor and nothing to the other players. Thus, investing publicly was equivalent to cooperating while investing privately was equivalent to defecting, or free-riding, in standard PGGs. Players 1 - 4 were assigned to groups ex-post (Horton et al. 2011) to create two conditions: the 'typical defector' condition (3 defectors, 1 cooperator) and the 'atypical defector' condition (3 cooperators, 1 defector).

Player 5 observed the decisions of Players 1 - 4, either in the typical defector condition (n = 102) or the atypical defector condition (n = 108). Player 5 was allocated $1.05 and could choose whether to pay a fixed cost ($0.05) to reduce the earnings of any of the other players by $0.15. Player 5 could punish one, two, three or all four of the PGG players for the same fixed cost of $0.05; thus, the increasing costs associated with punishing more than one player were removed in this game.

Subsequent to the punishment decision, Player 5 rated each PGG player on a seven point scale as to how much they would like to play a subsequent investment game with that player (similar to Irwin & Horne 2013; Parks & Stone 2010). This

answer provided a measure of social rejection. The majority of ratings were either one or seven (proportion = 0.68 ± 0.2) so we re-categorised ratings into a binary variable for analysis. Ratings less than four were set as 1 (indicating desire to avoid the player in question) and ratings of four or more were set as 0 (indicating indifference, or preference for the player in question).

Data were analysed using R version 3.02 (R Development Core Team 2011). Using two Generalized Linear Mixed Models (GLMMs), we measured the probability that a player would be (i) punished and (ii) socially rejected by Player 5 according to how they behaved (cooperator / defector) and whether or not the behaviour violated the descriptive norm in that setting. We additionally controlled for the effects of age and gender on Player 5's propensity to punish. We employed a multi-model inference approach (Grueber et al. 2011). Input variables were standardized (Gelman 2008). We estimated the importance and model-averaged coefficients of parameters using a set of models with the highest support (within 2AICc units of the top model; Burnham & Anderson 2004). We only present the parameter estimates from the top models (see General methods for further details).

7.5 Results

In general, typical and atypical behaviours were equally likely to be punished (proportion of typical behaviour punished = 0.17 ± 0.02; versus atypical = 0.22 ± 0.04; Table 7.1). In addition, defectors were just as likely to be punished whether their behaviour was typical (0.36 ± 0.03) or atypical (0.36 ± 0.05; Table 7.1; Figure 7.1). Similarly, cooperators were rarely punished, regardless of whether their behaviour was typical (0.02 ± 0.01) or atypical (0.01 ± 0.01; Table 7.1; Figure 7.1). Cooperators were never singled out for costly punishment and only faced punishment when all members of their group were also punished (on 3 occasions). Furthermore, when Player 5 invested to punish defectors, they always punished all defectors in the group rather than singling one individual out for punishment. Punishment was linked to gender, with male players being more likely to punish than females (proportion of individuals that were punished by males = 0.22 ± 0.02; versus females = 0.12 ± 0.02; Table 7.1).

The results for social rejection mirrored the punishment investment decisions above: cooperative individuals were preferred as partners over defectors for a hypothetical subsequent PGG, regardless of whether cooperative behaviour was typical or atypical (proportion defectors rejected *typical* = 0.84 ± 0.03; *atypical* = 0.80 ± 0.05). Although, players appear to reject atypical cooperators slightly more often than typical cooperators, the confidence intervals for the interaction term just crossed zero, meaning that the evidence for this effect is weak (cooperators rejected *typical* = 0.3 ± 0.01; *atypical* = 0.5 ± 0.02; Table 7.2; Figure 7.S1).

Parameter	Estimate	Unconditional SE	Confidence Interval	Relative Importance
Intercept	-4.35	0.56	(-5.45, -3.25)	
PGG decision (cooperate / defect)	5.98	1.02	(3.96, 8.00)	1.00
Player 5 gender (female / male)	2.19	0.49	(1.22, 3.16)	1.00
Player 5 age	-0.25	0.42	(-1.08, -0.59)	0.30

Table 7.1 Estimates, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models explaining whether PGG players were punished by Player 5.

Parameter	Estimate	Unconditional SE	Confidence Interval	Relative Importance
Intercept	-1.31	0.49	(-2.27, -0.34)	
PGG decision (cooperate / defect)	9.45	0.99	(7.51, 11.39)	1
Violated the descriptive norm (no / yes)	0.56	0.81	(-1.02, 2.14)	0.8
Violated the descriptive norm x PGG decision	-2.21	1.59	(-5.32, 0.90)	0.8
Player 5 gender (female / male)	0.74	0.56	(-0.36, 1.82)	0.62

Table 7.2 Estimates, unconditional standard errors, confidence intervals and relative importance for parameters included in the top models explaining whether PGG players were socially rejected by Player 5.

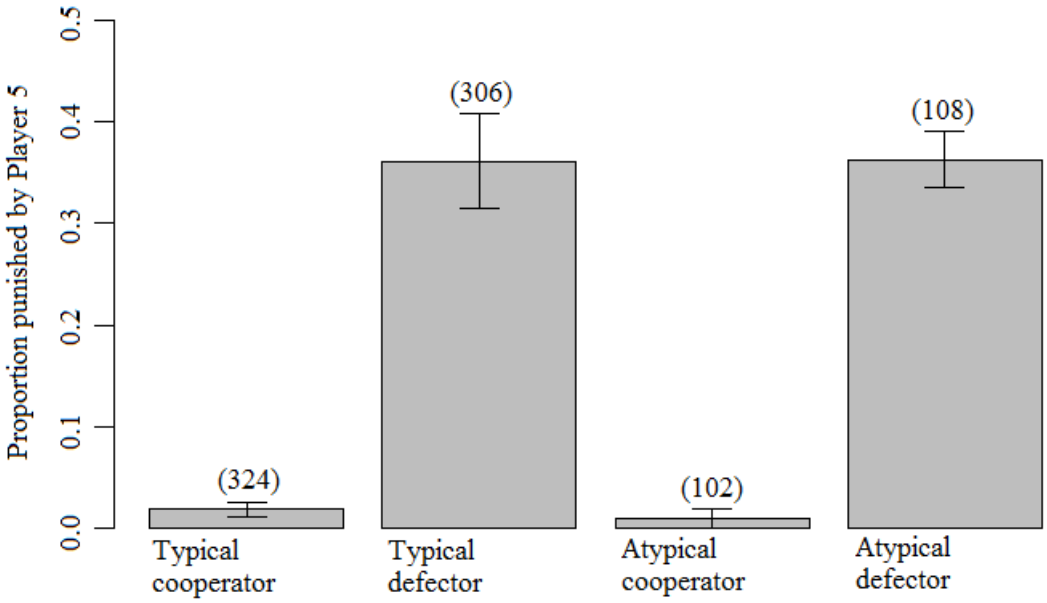


Figure 7.1 The proportion of PGG players who were punished by Player 5, according to their PGG decision and whether this violated the descriptive norm. Sample sizes for each condition are indicated in parentheses. Error bars show standard errors.

7.6 Discussion

Previous studies have suggested that punishment might be proximately driven by the desire to harm individuals that violate descriptive norms. However, these studies have typically not controlled for the possibility that paying to harm descriptive norm violators is less costly than paying to harm conformers, because the costs of punishing typically scale with the number of individuals that are punished (Fehr & Fischbacher 2004b; Horne 2009; Irwin & Horne 2013). Here, we removed this scaling effect of punishment by allowing individuals to pay a fixed cost to punish any or all of the PGG players. Under these conditions, individuals directed almost all punishment towards defectors regardless of whether defecting was the descriptive norm. These results contradict the prediction that defectors are less likely to be punished when they are typical (Fehr & Fischbacher 2004b) and suggest that defectors are probably viewed negatively regardless of their prevalence in the population. In other studies, rare defectors may receive more punishment than common defectors because this is less costly

to the punisher. It is possible that defectors were punished regardless of their prevalence because individuals did not make punishment decisions based on the events in the game but instead on a pre-existing perception of defection as a descriptive norm violation formed from their experience in the 'real world'. However, previous studies in the same cultural group (US-based subjects) have shown that individuals' behaviour is sensitive to similar descriptive norm manipulations that occur within the confines of the game setting (Irwin & Horne 2013; Parks & Stone 2010).

We found very little evidence for antisocial punishment in this setting, even when cooperators were in the minority. This contradicts previous findings, which have shown that excessively generous individuals are singled out for punishment, even though their behaviour ostensibly benefits the individuals who punish them (Irwin & Horne 2013). The rarity of antisocial punishment in our current study may be because many of the motives proposed to underpin antisocial punishment were absent in our setting. Most previous studies of antisocial punishment have shown that it comes from individuals within the group, rather than third-parties, suggesting that antisocial punishment reflects competition for status within groups (Sylwester et al. 2013). For example, antisocial punishment might occur in retaliation for punishment received (or expected to be received) from cooperators (Sylwester et al. 2013; Herrmann et al. 2008). Alternatively, since individuals are often chosen as partners based on their cooperativeness relative to others (Roberts 1998; Barclay & Willer 2007; Sylwester & Roberts 2010), defectors might punish cooperators because cooperators 'raise the bar', making defectors look bad in comparison (Herrmann et al. 2008; Minson & Monin 2012). In the absence of these motives, we found no evidence to suggest that descriptive norm deviants were more likely to be punished by third-parties. Our measures of social rejection, however, did hint that atypical cooperators were slightly less likely to be preferred for subsequent hypothetical interactions, when Player 5 would then be in the group with this individual. This tendency, although weak, supports previous work showing that excessively helpful, cooperative or moralistic individuals might be viewed negatively rather than positively by others in their social group (Monin et al. 2008; Parks & Stone 2010; Irwin & Horne 2013; Raihani 2014).

To summarize, third-party punishers targeted defectors, rather than descriptive norm violators in this setting. We suggest that decreased punishment of defectors when common might reflect the increased cost of punishing. Although, atypical cooperators were infrequently punished in this setting, they were slightly less preferred for subsequent interactions. Thus, the lack of antisocial third-party punishment in our setting might reflect the fact that punishers were not in competition for status with cooperators (Sylwester et al. 2013). Punishment of cooperative descriptive norm violators might be more common from fellow group members, rather than third-parties.

7.7 Supplementary materials

7.7.1 Supplementary methods

All subjects were paid a show-up fee of $0.30 on top of a bonus based on their payoff in the game. All Players (1 – 5) were required to answer nine comprehension questions. Players were required to answer all comprehension questions correctly to take part in the game.

Player 5's initial endowment was chosen so as to exceed the largest possible payoff of any of the PGG players to rule out disadvantageous inequality aversion as a motive for punishing defectors (e.g. Johnson et al. 2009; Raihani & McAuliffe 2012).

At the end of the experiment players were asked on a seven point scale (1 = not at all, to 7 = very much) how confident they were that the other players in their group were real people. It is important that players were confident that they were playing real people because punishment decisions have been shown to be different when people play human players compared to computer players (Sanfey et al. 2003). We repeated all analyses excluding players that gave an answer of less than four (proportion answering $< 4 = 0.46 \pm 0.02$) to the question above but found that this did not change the key findings of the study. Although the proportion of players that gave an answer of less than four was relatively high, it is likely that by asking this question we increased their suspicion that they weren't really playing real people.

All GLMMs produced had a binomial error distribution and logit link and were fitted using the lme4 package (Bates et al. 2011). GLMMs allow repeated measures to be fitted as random terms, thus controlling for their effects on the distribution of the data. We ran two models with our data. In model (i), we asked whether each of the Players 1 - 4 (n = 840) was punished by Player 5. The response term was set as '1' if the player was punished and '0' if the player was not punished. In model (ii), we asked whether the player was socially rejected by Player 5. Again, the response term was set as '1' if a player was rejected (preference score of < 4) and '0' if a player was not rejected (preference score of > 4). For both analyses, we included the following explanatory variables in the model: 'PGG decision' (cooperate / free-ride), 'violated the descriptive norm' (no / yes) and the two-way interaction between these variables. We also controlled for players' age and gender (male / female). For each model, 'Player 5 ID' was included as the random term.

7.7.2 Matching of subjects

As described in the General Methods section, because we did not know how many players would defect and how many would cooperate, it was not always possible to uniquely match players. In order to make up 104 groups which contained 3 defectors and 1 cooperator and 102 groups which contained 3 cooperators and 1 defector (as were presented to Player 5s), we would have required 426 defectors and 414 cooperators. In fact, 311 player 1 – 4s defected and 529 player 1 – 4s cooperated. This meant that 115 cooperting player 1 – 4s (11 % of subjects) had to be matched with 2 different Player 5s.

7.7.3 Supplementary results

An equal proportion of Player 5s invested in costly punishment in both the typical defector (0.36 ± 0.05) and the atypical defector treatment (0.36 ± 0.05). In both treatments, costly punishment was focussed on defectors rather than cooperators. Thus, antisocial punishment was uncommon in this setting (proportion of defectors punished = ± 0.36 ± 0.02; proportion cooperators punished = 0.02 ± 0.01).

Treatment	Player	PGG decision	Payoff	Descriptive Norm violated
Atypical defector	Player 1	Cooperate	$0.60	No
	Player 2	Cooperate	$0.60	No
	Player 3	Cooperate	$0.60	No
	Player 4	Defect	$0.90	Yes
	Player 5	NA	$1.05	NA
Atypical cooperator	Player 1	Defect	$0.50	No
	Player 2	Defect	$0.50	No
	Player 3	Defect	$0.50	No
	Player 4	Cooperate	$0.20	Yes
	Player 5	NA	$1.05	NA

Table 7.S1 The public goods game (PGG) decision, payoff received and whether or not this PG game decision violated the descriptive norm for Players 1 – 5 in the atypical cooperator and atypical defector treatment.

Parameter	Players 1 - 4	Player 5
Age	Mean = 29.36 ± 0.3	Mean = 30.14 ± 0.67
	Median = 27	Median = 27.5
	IQR = 24 – 33	IQR = 23 – 33.25
	Range = 18 - 66	Range = 18 -68
	Prefer not to disclose = 6	Prefer not to disclose = 2
Education level (n)	Some High School = 8	Some High School = 1
	High School Graduate = 76	High School Graduate = 14
	Some College, no degree = 282	Some College, no degree = 64
	Associates Degree = 64	Associates Degree = 20
	Bachelor's Degree = 322	Bachelor's Degree = 79
	Graduate Degree = 82	Graduate Degree = 32
	Prefer not to disclose = 6	Prefer not to disclose = 0
Gender (n)	Females = 313	Females = 67
	Males = 521	Males = 143
	Prefer not to disclose = 6	Prefer not to disclose = 0
Annual income (n)	Less than $12,500 = 79	Less than $12,500 = 20
	$12,500 - $24,999 = 105	$12,500 - $24,999 = 37
	$25,000 - $37,499 = 142	$25,000 - $37,499 = 24
	$37,500 - $49,999 = 101	$37,500 - $49,999 = 33
	$50,000 - $62,499 = 101	$50,000 - $62,499 = 30
	$62,500 - $74,999 = 88	$62,500 - $74,999 = 15
	$75,000 - $87,499 = 42	$75,000 - $87,499 = 6
	$87,500 - $99,999 = 39	$87,500 - $99,999 = 16
	$100,00 or more = 86	$100,00 or more = 21
	Prefer not to disclose = 57	Prefer not to disclose = 8

Table 7.S2 Information on mean and median values (where appropriate) and sample sizes for demographic information for Players 1 – 4 and Player 5.

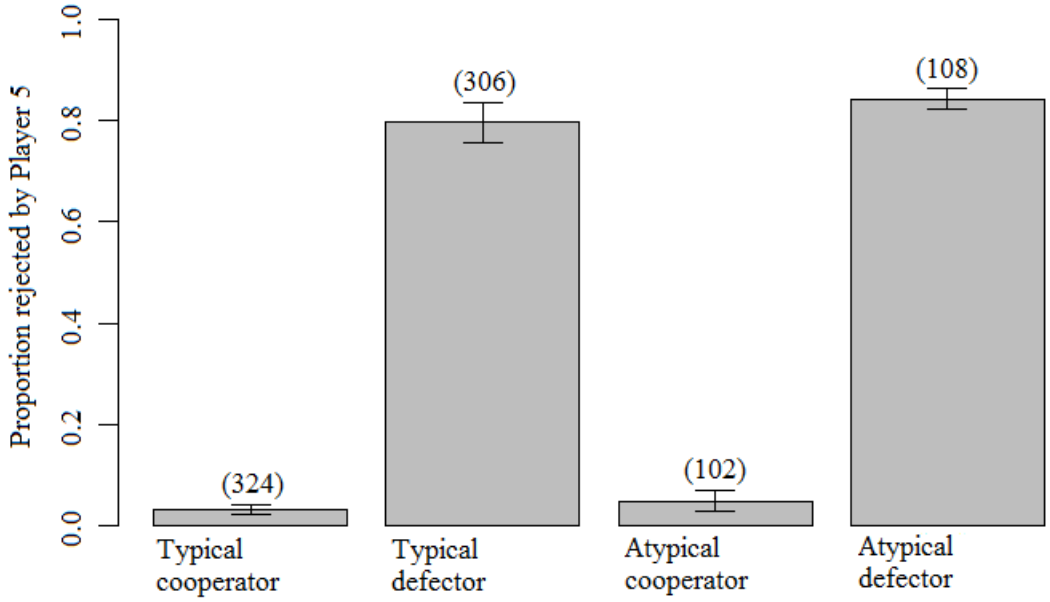


Figure 7.S1 The proportion of PGG players who were socially rejected by Player 5, according to their PGG decision and whether this violated the descriptive norm. Sample sizes for each condition are indicated in parentheses. Error bars show standard errors.

Chapter 8
General Discussion

8.1 Overview

Cooperative behaviour often involves individuals making a short-term investment (Bshary & Bergmüller 2008). The question of how this investment is repaid in terms of lifetime fitness has captured the attention of researchers across a wide-range of disciplines including evolutionary biology, economics and psychology. A plethora of mechanisms have been enlisted as possible solutions to this problem. One such mechanism is punishment (Gardner & West 2004a; Henrich & Boyd 2001; Boyd et al. 2003; Gintis 2000; Fehr & Gächter 2000; Fehr & Gächter 2002; Gächter et al. 2008). Despite the attention that punishment has received, many questions remain unanswered regarding when punishment is most effective at promoting cooperation and what proximate motivations underpin punishment decisions. In this thesis, I investigated a number of these unanswered questions using experimental economic games.

8.2 When does punishment promote cooperation?

Recent work has suggested that punishment use is detrimental because it induces retaliation rather than cooperation (Dreber et al. 2008; Janssen & Bushman 2008; Nikiforakis 2008). However, previous studies have typically made the unrealistic assumption that players are equal in terms of the power with which they can punish (e.g. Dreber et al. 2008; Egas & Riedl 2008; Fehr & Gächter 2002). I investigated the effect of power asymmetries on cooperation and punishment in an iterated prisoner dilemma (IPD) game, both where cooperation was binary (Chapter 3) and where cooperation investment was variable (Chapter 4).

Based on previous work in animal models (Clutton-Brock & Parker 1995; Axelrod 1984; Raihani, Thornton, et al. 2012; Wang et al. 2010; Bshary et al. 2008; Raihani et al. 2010; Raihani, Grutter, et al. 2012), I predicted that power asymmetries may stabilize cooperation in humans by making punishment from strong players (i) more effective at eliciting cooperation from defecting partners and (ii) cheaper, by reducing the costs associated with targets retaliating. However, my findings did not support these predictions. In both the binary investment IPD game and the variable investment IPD game, punishment from

strong players did not elicit cooperation from previously defecting weak partners in either symmetric or asymmetric games. In addition, weak players readily retaliated against strong partners, in both the binary investment IPD game (Chapter 3) and in the variable investment IPD game (Chapter 4). These findings indicate that power asymmetries did not stabilize punishment in the IPD games I studied.

These findings seem to contradict the findings from empirical work on non-human animals. For example, in the cleaner fish system, female fish never punish or retaliate against males (the larger sex) and females that are punished by a male behave more cooperatively in future interactions with that male (Raihani et al. 2010; Raihani, Grutter, et al. 2012). I propose that the difference between the findings of these human experiments and the cleaner fish studies may be associated with the acceptance of the dominant individual's authority. In my experiments (Chapter 3 & 4), the allocation of players' roles (weak or strong) was random. For this reason players may not have accepted the authority of strong players in asymmetric games. For example, a recent study demonstrated that punishment is likely to be more effective at promoting cooperation when centralized authorities have been legitimately elected; rather than chosen at random (Baldassarri & Grossman 2011).This situation resembles most industrialised democratic societies where punishment is largely handed over to a legal authority overseen by an elected government (Baldassarri & Grossman 2011; Kümmerli 2011; Rustagi et al. 2010).

When punishment is not centralized, punishment might only occur if a critical mass of peers agree to participate in the punishing the cheat (Mathew & Boyd 2011). For example, the Turkana, a pastoral society in East Africa, sustain costly cooperation in combat by the collective punishment of cowards or deserters. Importantly, punishment only occurs if a critical mass of peers assemble to punish the cheat (Mathew & Boyd 2011). In addition, laboratory studies have demonstrated that players prefer to pay a tax for the punishment of cheats to be performed on the behalf of the group, rather than conduct punishment themselves (Traulsen et al. 2012). Punishment may be more likely to elicit cooperation rather

than retaliation when it is supported by the consensus, either because it is conducted by an elected central authority (Baldassarri & Grossman 2011) or is coordinated by members of the group 'voting with their feet' (Mathew & Boyd 2011; Traulsen et al. 2012). Future work should explore whether power asymmetries allow punishment to promote cooperation when individuals are elected to the dominant role based on their behaviour in a preliminary task. In addition, real life power asymmetries may often resemble the relationship between a centralised authority or a group of individuals agreeing to punish and a lone cheat, rather than between one strong and one weak individual. Therefore, the importance of power asymmetries on eliciting cooperation in these situations deserves further consideration. It is possible that the acceptance of authority is also culturally dependent. Hofstede (1984) characterized world cultures by various dimensions, one of which - 'power distance' - describes the strength of social hierarchy. Individuals in a society that exhibits a high power distance are expected to accept hierarchies in which everyone has a place without the need for justification. However, individuals in societies with low power distance are expected to seek an equal distribution of power (Hofstede 1984). According to power distance values taken from The Hofstede Centre website (www.geert-hofstede.com), the power distance in the UK, where my IPD experiments took place, is relatively low (power distance = 35) in comparison to the average power distance across all countries for which data is available (average power distance = 59). It is possible that we would have seen a greater effect of power asymmetries had my experiments taken place in a country with a higher power distance because players may have more readily accepted the strong player's dominant position. Future work, should explore how cultural differences affect the acceptance of punishment from dominant individuals.

In Chapter 3 & 4, being punished (even by a strong player) meant losing a relatively small amount of money. I suggest that this may explain why punishment was ineffective at promoting cooperation in these experiments but in cleaner fish, where punishment was more costly or even potentially fatal, punishment does promote cooperation from targets (Raihani et al. 2010). I suggest that the relatively small cost associated with being punished may also explain why in

these human experiments weak players punished and retaliated against strong partners in but in the cleaner fish system**,** *female fish never punish or retaliate against larger, dominant males* (Bshary et al. 2008; Raihani et al. 2010; Raihani, Grutter, et al. 2012). On the other hand, anthropologic studies of real-life suggest that most punishment comes in the form of ridicule, gossip and verbal reproach (Boehm 1993), all of which carry a relatively low cost to the target. I therefore suggest that future work should explore the effect of power asymmetries on cooperation and punishment with a range of fee-to-fine ratios, both higher and lower than those tested in Chapter 3 & 4.

Several studies have demonstrated that subtle manipulations in the context given in laboratory game instructions can have drastic effects on subjects' behaviour (Kagel & Roth 1995; Hertwig & Ortmann 2001; Bohnet & Cooter 2003). For example, defection rates have been shown to vary depending on whether a prisoners dilemma game is described to subjects as a "Community" or a "Wall Street" game (Ross & Ward 1996) and punishment rates vary depending on whether subjects are given the option to "punish" or to "assign" points" to other players (Gintis 2001). In order to avoid such framing effects, neutrally worded instructions have become a mainstream practice in behavioural experiments. I suggest that the use of neutral language in Chapters 3 and 4 may explain why weak and strong players did not behave as expected. For example, although players were aware of the different payoff consequences of actions performed by the two player types, weak and strong players were referred to as 'Type 1' and 'Type 2' respectively. The use of these neutral labels ('Type 1' and 'Type 2') may have not been salient enough to translate into dominant and subordinate social roles in players' minds. Future work should explore how players behave in similar experiments when they are explicitly told that they are playing the role of a subordinate or dominant individual. Increasing the saliency of the roles allocated to players may also be achieved by setting experimental games within a context in which people are already familiar with the different roles; for example the game could be framed as a 'workplace game' and players allocated the role of 'junior' or 'senior' workers (Nikiforakis et al. 2014). Moreover, experimental games could be performed in the presence of real life power asymmetries, for example between real life bosses and employees. Such an experiment would particularly interesting as studies have suggested that subjects bring context learned in everyday life into the lab (Levitt & List 2007).

Previous studies have suggested that the more efficient punishment is (i.e. the cheaper punishment is for the punisher, relative to the amount the target's payoff is reduced by), the more effective it is at promoting cooperation (Nikiforakis & Normann 2008; Egas & Riedl 2008; Falk et al. 2005; Vukov et al. 2013). My results supported these previous findings to some extent. Although, incurring punishment did not increase players' cooperation in the next round in either the binary investment IPD (Chapter 3) or the variable investment IPD experiment (Chapter 4), players were more cooperative in general when paired with a strong partner than a weak partner. This indicates that the mere threat of punishment (but not the use of punishment) from a strong player may have more effectively deterred cheating than the threat of punishment from a weak partner (i.e. more efficient punishment was a more effective threat). If the mere threat of punishment is sufficient to deter cheats, the costs associated with punishment will be lower because punishment will rarely need to be executed (Cant 2011). However, it is likely that the threat of punishment will quickly lose its power as a deterrent if it is not backed up with actual punishment in response to cheating. In fact, since the use of punishment may increase the target's sensitivity to the threat of future punishment (Ellis 2012), by only investigating the target's cooperation directly after punishment I may have missed some of the longer term deterrent effects associated with incurring punishment. For example, although I did not find a positive effect of punishment in the round directly after targets were punished (potentially because some players responded by defecting in retaliation), punishment may have had a deterrent effect for several subsequent rounds and potentially the rest of the game. Unfortunately, I was unable to test this hypothesis with the current set-up because analysing players' decisions in the rounds following punishment would have created an unmanageable number of confounding factors. This hypothesis could be tested in future experiments by incorporating a stooge player into the game that is constrained to play a filay strategy or by hiding punishment decisions such that cheats are not informed whether or not they were punished by their peers until all rounds of the game have been completed (Vyrastekova et al. 2008; Fudenberg & Pathak 2010).Nevertheless, some of my other findings were less supportive of the idea

that more efficient punishment is more effective at promoting cooperation. In the binary investment IPD game (Chapter 3), strong players were actually more likely to defect (rather than cooperate) after being punished by strong partners and both player types were more likely to retaliate after being punished by strong partners. I propose that the detrimental effect of punishment from strong players may be related to the inferences made by targets regarding the punisher's motives. Fehr & Rockenbach (2003) argued that punishment may be perceived as being morally illegitimate if it is associated with selfish or greedy (rather than altruistic) intentions. It is possible that punishment that improves the punisher's payoffs relative to those of the target (as punishment from strong players did in my experiments), may be interpreted as a competitive act and therefore perceived as morally illegitimate (Xiao 2013; Fehr & Rockenbach 2003; Raihani & Bshary 2015). Previous studies have proposed that punishment which is perceived to be morally illegitimate is unlikely to promote cooperation from targets (Xiao 2013; Fehr & Rockenbach 2003). A logical extension of this prediction might be that morally illegitimate punishment makes targets more likely to defect and to retaliate; as observed when strong players punished strong partners in the binary investment IPD experiment (Chapter 3). Future work could explore the moral assessment of punishment in different circumstances by collecting data on both players' behavioural responses as well as their subjective evaluations of punishment and the reasons why they decided to retaliate.

The fact that punishment did not promote cooperation from targets in either IPD experiment (Chapters 3 & 4), supports previous work which has suggested that in a two-player setting, conditional cooperation may sustain cooperation more effectively than punishment (Rand et al. 2009). Further support for this finding comes from the finding that in the binary IPD game, defecting players were more likely to switch to cooperate if their partner cooperated in the previous round. Moreover, in the variable investment IPD, players that chose a lower cooperation level than their partner were more likely to increase their cooperation in the following round than to decrease or choose same cooperation level again. This finding is indicative of the existence of 'give-as-good-as-you-get' strategies (Roberts & Sherratt 1998). Nevertheless, conditional cooperation may be less

effective in multi-player games because defecting in response to the defection of a group member will also harm the cooperative members of the group (Raihani, Thornton, et al. 2012). This suggests that whilst punishment did not promote cooperation in these two-player games, punishment from strong players may be more effective in a multi-player game (e.g. Przepiorka & Diekmann 2013). Indeed, many real-life social dilemmas, such as climate change or overfishing, more closely resemble multiplayer than two-player games (Raihani & Hart 2010). Future work should explore how power asymmetries effect punishment and cooperation in multiplayer games. Although, Nikforakis et al. (2010), incorporated power asymmetries into a multiplayer game, they did not test whether the effectiveness of punishment use was affected by power asymmetries.

A possible limitation to the IPD experiments reported in Chapters 3 & 4 is that there was not a treatment in which punishment was not possible. The inclusion of this control treatment would have allowed me to investigate how the availability of punishment affected cooperation levels in the games used. For example, even though punishment itself did not increase cooperation from targets in the next round, conditionally cooperative strategies may not have been so successful if not backed up by a threat of punishment. The possibility of punishment may have encouraged conditional cooperation by increasing players' expectations that their partner would also cooperate.

8.3 The proximate motivations underpinning punishment decision

Recent work has suggested that both second-party and third-party punishment is motivated by negative emotions (Fehr & Gächter 2002; Sanfey et al. 2003; Xiao & Houser 2005; Grimm & Mengel 2011; Wang et al. 2011; Fehr & Fischbacher 2004b; Jordan et al. 2014). However the source of these negative emotions is not well understood.

8.3.1 A desire for revenge versus desire for equality

Previous work has suggested that the negative emotions underpinning second-party punishment could arise from one of two (not mutually exclusive) sources (Raihani & McAuliffe 2012a): (i) a desire to inflict reciprocal harm on a cheating partner (also termed the desire for 'revenge'; McCullough et al. 2013) or (ii) a desire for equality (Raihani & McAuliffe 2012b; Johnson et al. 2009). My findings from Chapter 5 suggested that punishment is motivated by both a desire for revenge *and* a desire for equality. Support for the idea that punishment is motivated by a desire for revenge comes from the findings that players punished stealing partners regardless of whether stealing created disadvantageous inequality for the punisher and players punished when punishing was unable to restore equality. However, players' tendency to punish increased if stealing resulted in disadvantageous inequality for the punisher and when possible, punishers most often tailored their investment in punishment to create equal outcomes, suggesting that players were also motivated by a desire for equality.

Whilst my results in Chapter 5 suggested that punishment stems from both a desire for revenge and an aversion to inequality, previous work using a similar experimental set-up found that players only punished stealing partners when stealing resulted in disadvantageous inequality (Raihani & McAuliffe 2012b), suggesting that punishment was motivated by egalitarian preferences and not a desire for revenge. My findings in Chapter 6 suggest the different demographic sampling of the two studies could explain different findings of these two studies (whilst I only recruited subjects from the US in Chapter 5, Raihani & McAuliffe 2012b) recruited from both the US and India). I found that whilst for US-based subjects, the decision to punish was affected both by whether the partner stole and by whether the punisher experienced disadvantageous inequality, Indian subjects only punished stealing partners more than non-stealing partners when stealing resulted in disadvantageous inequality (as in Raihani & McAuliffe 2012b)

It has been proposed by (Brosnan 2006; Brosnan 2011; but see Chen & Santos 2006) that an aversion to inequality is most likely to evolve in species that regularly cooperate with non-kin. According to Brosnan (2011), the ability to detect and respond to inequality may be beneficial in the social domain because it

can encourage individuals to avoid unfair partners (see also Baumard et al. 2013). Following, Price et al. (2002), I propose that inequality aversion may also be of benefit in social contexts because it allows cooperators to recognise and therefore act to remove the fitness-differentials between cooperators and cheats using punishment. Previous studies profess to have demonstrated inequality aversion in several non-human primate species (e.g. Brosnan 2006; Brosnan et al. 2010; Neiworth et al. 2009), all of which are species that are known to cooperate with unrelated group members; however, such claims have not gone undisputed (e.g. Silberberg et al. 2009; Jensen et al. 2007; Bräuer et al. 2006). Nevertheless, since evidence for punishment in non-human species is weak (Jensen et al. 2007a; Jensen et al. 2007b; Raihani, Thornton, et al. 2012), these examples of inequality aversion do little to inform our understanding of the relationship between inequality aversion and punishment (Raihani & McAuliffe 2012a). Since cleaner fish regularly cooperate with and punish non-kin during joint inspections, they provide a valuable non-human model to study inequality aversion in the context of punishment (Raihani & McAuliffe 2012a). Using the cleaner fish system, Raihani, Pinto, et al. (2012) showed that male punishment of cheating females does not require the male to compare his own payoffs with the females. Instead, males appear to take a cognitive short-cut by punishing in response to the client's sudden departure, an indirect signal that the female cheated. Moreover, a follow up study showed that cleaner fish are insensitive to outcome inequity when performing an effortful task (providing tactile stimulation to a model client) in return for food rewards (Raihani, McAuliffe, et al. 2012). This raises the question of why some species that cooperate with non-kin (i.e. humans and possibly primates) are apparently inequality averse whereas cleaner fish are not. One possible reason why humans are inequality averse whereas cleaner fish are not is that, cleaner fish's cooperative interactions are mostly limited to joint inspections of clients; humans however, cooperate in a vast number of different and often unique situations. Therefore, while punishing in response to the clients departure may be sufficient to restore 'fair' outcomes for cleaner fish (Raihani, Pinto, et al. 2012), humans may require a more versatile psychological mechanism (i.e. inequality aversion) to consistently arrive at fair payoff distributions.

Punishment might be perceived to be ‘unfair’ (and therefore less effective) if it increased the punishers payoff relative to the targets. Since cleaner fish do not appear to have fairness preferences *per se* (Raihani, McAuliffe, et al. 2012), how females respond to being punished by a male is not likely to be affected by how ‘fair’ she perceives the male to be. Although speculative, this may explain why female cleaner fish respond to punishment from male fish with increased cooperation (Bshary et al. 2008; Raihani et al. 2010; Raihani, Grutter, et al. 2012), whereas punishment from strong partners in my human experiments provoked retaliation from targets (Chapter 3 & 4).

I believe that the main limitation of Chapter 5 & 6 was that with the experimental set-up used, I was unable to test if punishment use was affected by whether or not outcome inequality was a result of the intentional actions of the target. Although, previous studies have shown that individuals will punish in response to unequal outcomes created at random or unintentionally (Cushman et al. 2009; Dawes et al. 2007; Falk et al. 2008; Houser & Xiao 2010; Kagel et al. 1996; Yu et al. 2014), this previous work did could not test whether players tailored their punishment investment to create equal outcomes under these circumstances. Future work should explore this possibility.

8.3.2 Dislike of descriptive norm deviants

Previous studies have suggested that third-party punishment is motivated by the violation of broadly recognized group norms (i.e. descriptive norms), rather than simply by a personal aversion to cheats or disadvantageous inequality (e.g. Fehr & Fischbacher 2004; Carpenter & Matthews 2012). For example, Irwin & Horne (2013) showed that individuals punished descriptive norm violators in a public goods game regardless of whether the target's behaviour was beneficial or detrimental to the group. It is possible that descriptive norm deviants are targeted for punishment because punishing the 'odd one out' is the cheapest strategy when the cost of punishing increases with the number of targets. However, I found that when the costs associated with punishment did not increase according to the number of players punished, defectors were equally likely to be punished when defection was common as when defection was rare (Chapter 7). This finding

suggests that in other studies, rare defectors may receive more punishment than common defectors because this is less costly to the punisher. Nevertheless, there is an important distinction between my study (Chapter 7) and that of Irwin & Horne (2013) which might also explain why I found that defectors were punished, irrespective of whether they were rare or common. In my study, the punisher was a third-party and therefore not part of the initial public goods game. By contrast, in the Irwin & Horne study (2013) punishers were part of the same group as their targets. Thus, in the Irwin & Horne (2013) study, defectors may have punished rare cooperators because cooperators 'raised the bar', making defectors look bad in comparison (e.g. as suggested by Herrmann et al. 2008; Minson & Monin 2012). Data from donations made to online fundraising pages has shown that people are sensitive to how their donation compares to those made by other donors, and will opt for anonymity when making very generous donations (Raihani 2014). This finding also suggests that individuals expect to be evaluated negatively when their contribution has the possibility to make the contributions of others look stingy.

8.3.3 Competitive motives and antisocial punishment

In all of my experiments (Chapter 3-7) I observed some level of punishment aimed at cooperative or non-stealing individuals, commonly referred to as 'antisocial punishment' (Herrmann et al. 2008; Ellingsen et al. 2012; Gächter & Herrmann 2009) or 'spite' (Abbink & Sadrieh 2009). Previous studies have suggested that antisocial punishment may be motivated by a desire to improve social status relative to others (Sylwester & Roberts 2013; Prediger et al. 2014; Bryson et al. 2014). My findings generally support this hypothesis. Firstly, previous work has suggested that the benefits of acquiring higher social status might be higher when the rule of law is weak and resources are scarce (for which GDP might provide a reasonable proxy; Sylwester et al. 2013; Prediger et al. 2014; Bryson et al. 2014). I would therefore predict that if antisocial punishment is motivated by competition for status then subjects from countries with weaker rule of law and lower GDP would be more likely to invest in antisocial punishment. Consequently, I would expect there to be more antisocial punishment

from India-based subjects than US-based subjects because GDP is lower and rule of the law is weaker in India compared to the US. In Chapter 6, I showed that India-based subjects were more likely to invest in antisocial punishment than US-based subjects. My findings therefore support this prediction. This is also consistent with another cross-cultural study which found that there was more antisocial punishment in societies with weak rule of law (Gächter et al. 2005). Second, if antisocial punishment is aimed at maximising relative payoffs I would expect that players would be more likely to invest in antisocial punishment if punishment improved the punisher's payoffs relative to those of the target (hereafter 'efficient punishment') than if punishment is equally costly to the punisher and the target (hereafter 'inefficient punishment'). This prediction has been supported by previous studies (Falk et al. 2005; but see Egas & Riedl 2008) and my findings partly supported this prediction. In the IPD experiments reported in Chapters 3 & 4, players were more likely to punish cooperative partners if they had access to efficient punishment than if they only had access to inefficient punishment. Nevertheless, in the experiment reported in Chapter 5, antisocial punishment was equally common when punishment was efficient as when it was inefficient. It is possible that I did not observe more antisocial punishment from players with efficient (rather than inefficient) punishment in Chapter 5 because in this experiment, punishers interacting with non-stealing partners were already better off than the target in all but one treatment, where players' payoffs were equal. However, it remains unclear why antisocial punishment was equally likely from efficient and inefficient punishers when players' outcomes were equal.

Nevertheless, alternative explanations for antisocial punishment exist. Previous studies have suggested that antisocial punishment might sometimes reflect retaliation by cheats who were punished by cooperative partners, or a pre-emptive strike against expected punishment in subsequent rounds (Cinyabuguma et al. 2006; Nikiforakis & Normann 2008; Herrmann et al. 2008; Sylwester et al. 2013; Raihani & Bshary 2015). This idea is supported by data from Chapter 3 and Chapter 4 where players in IPD games were more likely to antisocially punish their partner if they were punished in the previous round. This however, cannot explain all the antisocial punishment in Chapter 3 and 4, some of which took

place even when the punisher was not punished in the previous round. In addition, retaliation cannot explain the antisocial punishment I observed in one-shot games (Chapter 5, 6, 7). Although my findings best support the hypothesis that antisocial punishment is motivated by a desire to improve relative social status, antisocial punishment may be underpin by several different motives. Further dedicated experimentation is required to disentangle the relative importance of these possible motivations.

I used the online labour market Amazon Mechanical Turk (MTurk) to recruit subjects for the experiments reported in Chapter 5, 6 and 7. Online experimentation allows data to be collected internationally and enables data to be collected from a large number of individuals at a relatively low cost and within a relatively short amount of time. Nevertheless, the reliability of data collected via MTurk has been subject to debate. It has been argued that experimenters using MTurk relinquish a degree of control because they cannot be certain that subjects complete the task alone or if subjects are distracted by simultaneously performing other tasks (e.g. instant messaging or watching television; Bartneck, Duenser, Moltchanova, & Zawieska, 2015). However, in a survey of MTurk workers, most reported that they did complete tasks alone and were not engaged in other activities (Chandler et al. 2014). Nevertheless, in Chapter 5, 6 and 7, attention checks built in as comprehension questions were used to screen out subjects who either did not pay attention to or did not understand the task (Goodman, Cryder, & Cheema, 2013). Another common criticism of MTurk centres on the smaller stakes typically used in MTurk experiments in comparison to laboratory studies and the possibility that this may bias subjects behavior (Raihani, Mace, & Lamba, 2013). However, since the effective hourly rates of MTurk subjects who took part in the experiments reported in this thesis (around $19 per hour) were comparable with those in laboratory studies, I do not believe that this is a major concern in this case. Previous work has demonstrated that the MTurk subject base has become increasingly experienced with common behavioural experiments over time and that behaviour in some tasks has been shown to vary with the experience of subjects (Rand, Greene, & Nowak, 2012). However, alternative studies have found no systematic differences in the behaviour of experienced versus naive

subjects (Raihani & Bshary, in review) and that subjects behaviour is remarkably consistent across different cooperation games conducted on MTurk as well as in self-reports of real-life measures of cooperative tendency (Peysakhovich, Nowak, & Rand, 2014). In addition, several behavioural experiments using MTurk have reliably replicated findings originally obtained in a laboratory setting (Horton, Rand, & Zeckhauser, 2011; Rand, 2012; Suri & Watts, 2011). Taken together, therefore, I believe that the results obtained using MTurk in Chapter 5, 6 and 7 should be as reliable as those that could be obtained using laboratory methods.

8.4 Future work

A major limitation of the experiments reported in Chapter 5, 6 and 7 was the fact that the game only lasted one round, meaning that it was not actually possible for any deterrent function of punishment to be realised. It has been argued that since we evolved in a context where one-shot or anonymous interactions were rare (Delton et al. 2011), our evolved psychology is likely to invoke responses that are attuned to these conditions even one-shot lab settings (Ben-Ner & Putterman 2000; Burnham & Johnson 2005; Cosmides & Tooby 1989; Delton et al. 2011; Hagen & Hammerstein 2006; Hoffman et al. 1998; Johnson et al. 2003; Tooby et al. 2006). However, other studies have demonstrated that when players are given the time to consider their decisions, they are more likely to respond in a way that maximizes their payoff in their current one-shot setting (Grimm & Mengel 2011; Smith & Silberberg 2010; Sutter et al. 2003; Rand et al. 2012; Rand et al. 2014). It is therefore possible that players may have behaved differently in these studies if the game was repeated for multiple rounds (Grimm & Mengel 2011; Smith & Silberberg 2010; Sutter et al. 2003) before making their punishment decision. Future work should explore this possibility by using similar experimental setups as reported in Chapter 5, 6 and 7 but allowing players to interact for several rounds.

Furthermore, the experimental games reported in this thesis, allowed a very limited set of choices to players. For example, in the punishment stage of all the experiments, players could only choose between punishing or not punishing their partner. In reality however, there is likely to be the option to reward as well as to

punish partners. Although limiting the choices available to subjects simplified the analysis and interpretation of data produced from these experiments, providing a richer choice set could yield valuable insights. It is likely that dominant individuals will hold more resources and thus be able to grant larger rewards as well as inflict harsher punishment than subordinates. Thus, future work should explore the effect of asymmetries in players' ability to reward, as well as punish, on cooperation. The experiments reported in Chapter 5 and 6 could also be extended by allowing P2 to reward as well as to steal from P1 and allowing P1 to reward as well as to punish P2. It would be interesting to explore whether P1s who experienced disadvantageous inequality after being rewarded by P2 still punished P2 to create equal outcomes. It would also be valuable to test whether P1s who experienced advantageous inequality would reward P2s to create equal outcomes.

Along with the results from previous studies (Wu et al. 2009; Gächter et al. 2005; Balliet & Lange 2013; Herrmann et al. 2008; Henrich et al. 2006), the cross-cultural variation in the motivations underpinning punishment found in Chapter 6, highlight the importance of studying behaviours in a more diverse demographic sample than the typical western, educated, industrialised, rich and democratic (WEIRD; Henrich et al. 2010) samples used in the majority of behavioural experiments (Buhrmester et al. 2011). In addition, previous cross-cultural studies have typically sampled from one population per culture and have therefore confounded cultural and demographic differences (e.g. age structure and social network size) between populations (Lamba & Mace 2011). Future cross-cultural work should collect data from multiple different populations within each culture taking into account the demographics of sampled populations.

Experimental games studying cooperation typically assume that players have to interact with a randomly allocated partner and they have no scope for partner choice. However, in reality individuals may have opportunities for interactions elsewhere ('outside options'; Cant 2011; Raihani et al. 2012), and so they may be able to choose at which point to end an interaction in order to interact with an alternative partner. Outside options can also influence the outcome of cooperative

interactions, since players are expected to choose the most cooperative of their prospective partners to interact with (Roberts 1998; Barclay & Willer 2007; Noë & Hammerstein 1995; McNamara et al. 2008; Sylwester & Roberts 2010; Sylwester & Roberts 2013). However, individuals may vary in their outside options, generating asymmetries between players: some players may have several prospective interaction partners to choose from while others have relatively few. It is expected that in cooperative interactions, individuals with more outside options will assess the behaviour of potential partners and choose to interact with the most cooperative individuals. On the other hand, those with fewer outside options may invest in their reputation by behaving cooperatively in order to attract choosy partners (Barclay & Willer 2007; McNamara et al. 2008; Roberts 1998; Noë & Hammerstein 1995). It is therefore expected that where there are asymmetries in outside options, players with fewer outside options will be more cooperative than those with many alternative partners to choose from.

If punishment is an option it is likely that players with more outside options will only attempt to discipline a cheating partner up to a point before ending the interaction and choosing an alternative partner. Therefore, it is predicted that player's with fewer outside options will be more receptive to punishment and will be less likely to retaliate against punishment from players with more outside options. It is possible that this will make punishment more effective at eliciting cooperation. Previous work on cleaner fish supports this idea. Whilst clients who are forced to repeatedly interact with the same cleaner fish (due to small home ranges) punish cheating cleaners with aggressive chasing (Bshary & Grutter 2002), clients with access to several cleaners (due to large home ranges) rarely punish cheats and instead choose another cleaner fish (Bshary & Schäffer 2002).

Whether players prefer to punish cheats or to choose alternative partners is also likely to depend upon the variation in cooperativeness in the population (Leimar & Hammerstein 2010; McNamara et al. 2008). For example, if there is low variation in cooperativeness in the population then punishment might be preferred over switching because alternative partners are unlikely to be an improvement on the current partner. On the other hand, when population variation is higher,

switching might be preferred over punishing. Varying the availability of information about the cooperativeness of potential partners may also affect player's strategic decisions regarding punishment or partner switching. For example, when faced with a cheating partner, players may prefer to switch when they know that alternative partners are cooperative but may be more likely to punish if such information is not available.

Future work could test whether (i) players with fewer outside options are more cooperative when interacting with players with more outside options (ii) if asymmetries in outside options improve the effectiveness of punishment at eliciting cooperation and (iii) how information on the cooperativeness of potential partners affects players decisions to punish their current partner or switch to an alternative partner. These questions could be explored by incorporating asymmetries in outside options into an IPD game by varying the cost associated with switching partners.

Chapter 9

References

Abbink, K. & Sadrieh, A., 2009. The pleasure of being nasty. *Economics Letters*, 105(3), pp.306–308.

Alexander, R.D., 1987. *The Biology of Moral Systems*, A. de Gruyter.

Andersen, S. et al., 2011. Stakes matter in ultimatum games. *American Economic Review*, 101(7), pp.3427–3439.

Anderson, C.M. & Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, 54(1), pp.1–24.

Andreoni, J. & Gee, L.K., 2011. Gun For Hire: Does Delegated Enforcement Crowd out Peer Punishment in Giving to Public Goods?

Asch, S.E., 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), pp.1–70.

Axelrod, R., 1986. An Evolutionary Approach to Norms. *American Political Science Review*, 80(04), pp.1095–1111.

Axelrod, R.M., 1984. *The Evolution of Cooperation*, New York: Basic Books.

Balafoutas, L. & Nikiforakis, N., 2012. Norm enforcement in the city: A natural field experiment. *European Economic Review*, 56(8), pp.1773–1785.

Balafoutas, L., Nikiforakis, N. & Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45), pp.15924–15927.

Baldassarri, D. & Grossman, G., 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27), pp.11023–7.

Balliet, D. & Lange, P.A.M. Van, 2013. Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science*, 8, pp.363–379.

Balliet, D., Mulder, L.B. & Van Lange, P.A.M., 2011. Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin*, 137, pp.594–615.

Barclay, P., 2004. Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution and Human Behavior*, 25(4), pp.209–220.

Barclay, P. & Willer, R., 2007. Partner choice creates competitive altruism in humans. *Proceedings. Biological sciences / The Royal Society*, 274, pp.749–753.

Barrett, L. et al., 2000. Female baboons do not raise the stakes but they give as good as they get. *Animal behaviour*, 59(4), pp.763–770.

Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PloS One, 10*(4).

Bates, D., Maechler, M. & Bolker, B., 2011. lme4: Linear mixed-effects models using S4 classes. *Comprehensive R Archive Network*.

Baumard, N., André, J.-B. & Sperber, D., 2013. A mutualistic approach to morality: the evolution of fairness by partner choice. *The Behavioral and brain sciences*, 36, pp.59–78.

Bell, A. V, Richerson, P.J. & McElreath, R., 2009. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proceedings of the National Academy of Sciences of the United States of America*, 106, pp.17671–17674.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp.289 – 300.

Ben-Ner, A. & Putterman, L., 2000. On some implications of evolutionary psychology for the study of preferences and institutions. *Journal of Economic Behavior & Organization*, 43(1), pp.91–99.

Bochet, O., Page, T. & Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 60(1), pp.11–26.

Boehm, C., 1993. Egalitarian Behavior and Reverse Dominance Hierarchy. *Current Anthropology*, 34(3), p.227.

Boehm, C., 1999. The natural selection of altruistic traits. *Human Nature*, 10(3), pp.205–252.

Bohnet, I. & Cooter, R.D., 2003. Expressive Law: Framing or Equilibrium Selection? *SSRN Electronic Journal*.

Bone, J. et al., 2015. The Effect of Power Asymmetries on Cooperation and Punishment in a Prisoner's Dilemma Game. *PloS one*, 10(1).

Bone, J., Silva, A.S. & Raihani, N.J., 2014. Defectors, not norm violators, are punished by third-parties. *Biology letters*, 10(7).

Bone, J.E. & Raihani, N.J., 2015. Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*.

Boone, J.L., 1992. Competition, Conflict, and The Development of Social Hierarchies. In E. A. Smith & B. Winterhalder, eds. *Ecology, Evolution and Social Behavior*. Aldine de Gruyter, pp. 301–337.

Botelho, A. et al., 2005. Social norms and social choice. *NIMA Working Papers*.

Bowles, S. & Gintis, H., 2004. The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology*, 65, pp.17–28.

Boyd, R. et al., 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6), pp.3531–5.

Boyd, R., Gintis, H. & Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science (New York, N.Y.)*, 328(5978), pp.617–20.

Boyd, R. & Richerson, P.J., 1982. Cultural transmission and the evolution of cooperative behavior. *Human Ecology*, 10, pp.325–351.

Boyd, R. & Richerson, P.J., 1990. Group selection among alternative evolutionarily stable strategies. *Journal of theoretical biology*, 145, pp.331–342.

Boyd, R. & Richerson, P.J., 1989. The evolution of indirect reciprocity. *Social Networks*, 11, pp.213–236.

Boyd, R. & Richerson, P.J., 1988. The evolution of reciprocity in sizable groups. *Journal of theoretical biology*, 132(3), pp.337–56.

Boyd, R. & Richerson, P.J., 2005. *The Origin and Evolution of Cultures*,

Brandt, H., Hauert, C. & Sigmund, K., 2003. Punishment and reputation in spatial public goods games. *Proceedings. Biological sciences / The Royal Society*, 270(1519), pp.1099–104.

Bräuer, J., Call, J. & Tomasello, M., 2006. Are apes really inequity averse? *Proceedings. Biological sciences / The Royal Society*, 273, pp.3123–3128.

Brosnan, S.F., 2011. A hypothesis of the co-evolution of cooperation and responses to inequity. *Frontiers in Neuroscience*.

Brosnan, S.F. et al., 2010. Mechanisms underlying responses to inequitable outcomes in chimpanzees, Pan troglodytes. *Animal Behaviour*, 79, pp.1229–1237.

Brosnan, S.F., 2006. Nonhuman species' reactions to inequity and their implications for fairness. *Social Justice Research*, 19, pp.153–185.

Bryson, J.J. et al., 2014. Understanding and Addressing Cultural Variation in Costly Antisocial Punishment. , pp.201–223.

Bshary, A. & Bshary, R., 2010. Self-serving punishment of a common enemy creates a public good in reef fishes. *Current biology : CB*, 20(22), pp.2032–5.

Bshary, R. et al., 2008. Pairs of cooperating cleaner fish provide better service quality than singletons. *Nature*, 455(7215), pp.964–6.

Bshary, R. & Bergmüller, R., 2008. Distinguishing four fundamental approaches to the evolution of helping. *Journal of evolutionary biology*, 21(2), pp.405–20.

Bshary, R. & Grutter, A.S., 2002. Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. *Animal Behaviour*, 63(3), pp.547–555.

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

Buckholtz, J.W. et al., 2008. The neural correlates of third-party punishment. *Neuron*, 60(5), pp.930–40.

Buhrmester, M., Kwang, T. & Gosling, S.D., 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), pp.3–5.

Burnham, K.P. & Anderson, D.R., 2004. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer.

Burnham, K.P. & Anderson, D.R., 2002. *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach, 2nd edn. Springer-Verlag, New York*,

Burnham, T.C. & Johnson, D.D.P., 2005. The biological and evolutionary logic of human cooperation. *Analyse & Kritik*, (27(2005)).

Camerer, C.F., 2003. *Behavioral Game Theory: Experiments in Strategic Interaction (The Roundtable Series in Behavioral Economics)*, Princeton University Press.

Cameron, L. et al., 2009. Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics*, 93(7-8), pp.843–851.

Cameron, L.A., 1999. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, 37(1), pp.47–59.

Cant, M.A., 2011. The role of threats in animal cooperation. *Proceedings. Biological sciences / The Royal Society*, 278(1703), pp.170–8.

Carlsmith, K.M., Darley, J.M. & Robinson, P.H., 2002. *Why do we punish? Deterrence and just deserts as motives for punishment.*,

Carpenter, J., Verhoogen, E. & Burks, S., 2005. The effect of stakes in distribution experiments. *Economics Letters*, 86(3), pp.393–398.

Carpenter, J.P., 2007. The demand for punishment. *Journal of Economic Behavior & Organization*, 62(4), pp.522–542.

Carpenter, J.P. & Matthews, P.H., 2012. Norm Enforcement: Anger, Indignation or Reciprocity? *Journal of the European Economic Association*, 10(3), pp.555–572.

Carpenter, J.P., Matthews, P.H. & Ong'ong'a, O., 2004. Why punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, 14, pp.407–429.

Chandler, J., Mueller, P. & Paolacci, G., 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1), pp.112–30.

Charness, G., Cobo-Reyes, R. & Jiménez, N., 2008. An investment game with third-party intervention. *Journal of Economic Behavior and Organization*, 68, pp.18–28.

Charness, G., Masclet, D. & Villeval, M.-C., 2010. Competitive Preferences and Status as an Incentive : Experimental Evidence. *Working Papers*.

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

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

Cialdini, R.B., Reno, R.R. & Kallgren, C.A., 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), pp.1015–1026.

Cinyabuguma, M., Page, T. & Putterman, L., 2006. Can second-order punishment deter perverse punishment? *Experimental Economics*, 9(3), pp.265–279.

Claidière, N. & Whiten, A., 2012. Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological Bulletin*, 138, pp.126–145.

Clutton-Brock, T.H. & Parker, G.A., 1995. Punishment in animal societies. *Nature*, 373(6511), pp.209–16.

Colman, A.M., 2006. The puzzle of cooperation. *Nature*, 440(7085), pp.744–745.

Cosmides, L. & Tooby, J., 1989. Evolutionary psychology and the generation of culture, part II: Case study: A computational theory of social exchange. *Ethology and Sociobiology*, 10(1–3), pp.51–97.

Croson, R.T.A., 2001. Feedback in voluntary contribution mechanisms: An experiment in team production. In *Research in Experimental Economics*. Emerald Group Publishing Limited, pp. 85–97.

Cushman, F. et al., 2009. Accidental outcomes guide punishment in a "trembling hand" game. *PloS one*, 4(8).

Darwin, C., 1859. Origin of Species. *Library*, 475, p.424.

Dawes, C.T. et al., 2007. Egalitarian motives in humans. *Nature*, 446(7137), pp.794–6.

Dehaene, S., Dupoux, E. & Mehler, J., 1990. Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of experimental psychology. Human perception and performance*, 16(3), pp.626–41.

Delton, A.W. et al., 2011. Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32), pp.13335–40.

Delton, A.W. et al., 2013. Merely opting out of a public good is moralized: an error management approach to cooperation. *Journal of personality and social psychology*, 105(4), pp.621–38.

Denant-Boemont, L., Masclet, D. & Noussair, C.N., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, 33(1), pp.145–167.

Dreber, A. et al., 2008. Winners don't punish. *Nature*, 452(7185), pp.348–51.

Dreber, A., Fudenberg, D. & Rand, D.G., 2011. Who Cooperates in Repeated Games? *SSRN Electronic Journal*.

Egas, M. & Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings. Biological sciences / The Royal Society*, 275(1637), pp.871–8.

Ellingsen, T. et al., 2012. Civic Capital in Two Cultures: The Nature of Cooperation in Romania and USA. *SSRN Electronic Journal*.

Ellis, A., 2012. *The Philosophy of Punishment*, Andrews UK Limited.

Engelmann, D. & Nikiforakis, N., 2012. In the long-run we are all dead: On the benefits of peer punishment in rich environments. *Working Papers*.

Erdfelder, E., Faul, F. & Buchner, A., 1996. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28(1), pp.1–11.

Ertan, A., Page, T. & Putterman, L., 2009. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), pp.495–511.

Falk, A., Fehr, E. & Fischbacher, U., 2005. Driving Forces Behind Informal Sanctions. *Econometrica*, 73(6), pp.2017–2030.

Falk, A., Fehr, E. & Fischbacher, U., 2008. Testing theories of fairness--Intentions matter. *Games and Economic Behavior*, 62(1), pp.287–303.

Falkinger, J. et al., 2000. A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review*, 90, pp.247–264.

Fehl, K. et al., 2012. I dare you to punish me-vendettas in games of cooperation. A. Szolnoki, ed. *PloS one*, 7(9).

Fehr, E. & Fischbacher, U., 2004a. Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4), pp.185–190.

Fehr, E. & Fischbacher, U., 2003. The nature of human altruism. *Nature*, 425(6960), pp.785–91.

Fehr, E. & Fischbacher, U., 2004b. Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), pp.63–87.

Fehr, E., Fischbacher, U. & Gächter, S., 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1), pp.1–25.

Fehr, E. & Gächter, S., 2002. Altruistic punishment in humans. *Nature*, 415(6868), pp.137–40.

Fehr, E. & Gächter, S., 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4), pp.980–994.

Fehr, E. & Henrich, J., 2003. Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism. *SSRN Electronic Journal*.

Fehr, E. & Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. *Nature*, 422(6928), pp.137–40.

Fehr, E. & Schmidt, K.M., 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3), pp.817–868.

Fershtman, C., Gneezy, U. & List, J.A., 2012. Equity Aversion: Social Norms and the Desire to Be Ahead. *American Economic Journal: Microeconomics*, 4(4), pp.131–44.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), pp.171–178.

Fischbacher, U., Gächter, S. & Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71, pp.397–404.

Fischbacher, U. & Gaechter, S., 2006. Heterogeneous Social Preferences and the Dynamics of Free Riding in Public Goods. *Social Science Research Network Working Paper Series*.

Forsythe, R. et al., 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), pp.347–369.

Foster, K.R., Wenseleers, T. & Ratnieks, F.L.W., 2006. Kin selection is the key to altruism. *Trends in ecology & evolution*, 21(2), pp.57–60.

Fowler, J.H., 2005. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19), pp.7047–9.

Frean, M., 1996. The evolution of degrees of cooperation. *Journal of theoretical biology*, 182(4), pp.549–59.

Fudenberg, D. & Pathak, P.A., 2010. Unobserved punishment supports cooperation. *Journal of Public Economics*, 94(1-2), pp.78–86.

Gächter, S. & Herrmann, B., 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1518), pp.791–806.

Gächter, S., Herrmann, B. & Thöni, C., 2005. Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, 28(06), pp.822–823.

Gächter, S., Herrmann, B. & Thöni, C., 2010. Culture and cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1553), pp.2651–61.

Gächter, S., Renner, E. & Sefton, M., 2008. The long-run benefits of punishment. *Science (New York, N.Y.)*, 322(5907), p.1510.

Gardner, A. & West, S., 2004a. Cooperation and Punishment, Especially in Humans. *The American Naturalist*, 164, pp.753–764.

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

Gelman, A., 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15), pp.2865–73.

Gibson, C.C. & Marks, S.A., 1995. Transforming rural hunters into conservationists: An assessment of community-based wildlife management programs in Africa. *World Development*, 23(6), pp.941–957.

Gintis, H. et al., 2003. Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24, pp.153–172.

Gintis, H., 2000. Strong reciprocity and human sociality. *Journal of theoretical biology*, 206, pp.169–179.

Gintis, H., 2001. The Contribution of Game Theory to Experimental Design in the Behavioral Sciences. *Behavioral and Brain Sciences*, 24, pp.411–412.

Goodman, J.K., Cryder, C.E. & Cheema, A., 2013. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3), pp.213–224.

Greif, A., 1993. Contract Enforceability and Economic Institutions in Early Trade: the Maghribi Traders' Coalition. *American Economic Review*, 83(3), pp.525–48.

Grimm, V. & Mengel, F., 2011. Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, 111(2), pp.113–115.

Grueber, C.E. et al., 2011. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology*, 24(4), pp.699–711.

Grutter, A., 1996. Parasite removal rates by the cleaner wrasse Labroides dimidiatus. *Marine Ecology Progress Series*, 130(Losey 1974), pp.61–70.

Grutter, A.S. & Bshary, R., 2003. Cleaner wrasse prefer client mucus: support for partner control mechanisms in cleaning interactions. *Proceedings of the Royal Society B Biological Sciences*, 270(Suppl 2), pp.S242–S244.

Gürerk, O., Irlenbusch, B. & Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. *Science*, 312(5770), pp.108–111.

Hagen, E.H. & Hammerstein, P., 2006. Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theoretical population biology*, 69(3), pp.339–48.

Hamilton, W.D., 1972. Altruism and Related Phenomena, Mainly in Social Insects. *Annual Review of Ecology and Systematics*, 3, pp.193–232.

Hamilton, W.D., 1975. Innate Social Aptitudes of Man: an Approach from Evolutionary Genetics. In *Biosocial Anthropology*. pp. 133–155.

Hamilton, W.D., 1964a. The genetical evolution of social behaviour. I. *Journal of theoretical biology*, 7, pp.1–16.

Hamilton, W.D., 1964b. The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7, pp.17–52.

Haney, C., Banks, C. & Zimbardo, P., 1973. Interpersonal dynamics in a simulated prison. , pp.69–97.

Hardy, C.L. & Van Vugt, M., 2006. Nice guys finish first: the competitive altruism hypothesis. *Personality and social psychology bulletin*, 32, pp.1402–1413.

Hauert, C. et al., 2008. Public goods with punishment and abstaining in finite and infinite populations. *Biological theory*, 3(2), pp.114–122.

Hauert, C. et al., 2007. Via freedom to coercion: the emergence of costly punishment. *Science (New York, N.Y.)*, 316(5833), pp.1905–7.

Henrich, J. et al., 2006. Costly punishment across human societies. *Science (New York, N.Y.)*, 312(5781), pp.1767–70.

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

Henrich, J. et al., 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *The Behavioral and brain sciences*, 28, pp.795–815; discussion 815–855.

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

Henrich, J., Heine, S.J. & Norenzayan, A., 2010. Most people are not WEIRD. *Nature*, 466(7302), p.29.

Herrmann, B., Thöni, C. & Gächter, S., 2008. Antisocial punishment across societies. *Science (New York, N.Y.)*, 319(5868), pp.1362–7.

Hertwig, R. & Ortmann, A., 2001. Experimental practices in economics: a methodological challenge for psychologists? *The Behavioral and brain sciences*, 24, pp.383–403; discussion 403–451.

Hilbe, C. & Sigmund, K., 2010. Incentives and opportunism: from the carrot to the stick. *Proceedings. Biological sciences / The Royal Society*, 277(1693), pp.2427–2433.

Hilbe, C. & Traulsen, A., 2012. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific reports*, 2, p.458.

Hinrichs, J. V., Yurko, D.S. & Hu, J., 1981. Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), pp.890–901.

Hobbes, T., 1651. The Leviathan. *Journal of the American Chemical Society*, 79, pp.5019–5023.

Hoffman, E., McCabe, K. a. & Smith, V.L., 1998. Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology. *Economic Inquiry*, 36(3), pp.335–352.

Hoffman, E., McCabe, K.A. & Smith, V.L., 1996. On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory*, 25(3), pp.289–301.

Hofstede, G., 1984. *Culture's Consequences: International Differences in Work-Related Values*, SAGE Publications.

Horne, C., 2009. *The Rewards of Punishment: A Relational Theory of Norm Enforcement*, Stanford University Press.

Horton, J.J., Rand, D.G. & Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3), pp.399–425.

Houser, D. et al., 2008. When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, 62(2), pp.509–532.

Houser, D. & Xiao, E., 2010. Inequality-seeking punishment. *Economics Letters*, 109(1), pp.20–23.

Hurvich, C.M. & Tsai, C.-L., 1993. A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection. *Journal of Time Series Analysis*, 14(3), pp.271–279.

Irwin, K. & Horne, C., 2013. A normative explanation of antisocial punishment. *Social Science Research*, 42(2), pp.562–570.

Janssen, M.A., Anderies, J.M. & Joshi, S.R., 2011. Coordination and cooperation in asymmetric commons dilemmas. *Experimental Economics*, 14(4), pp.547–566.

Janssen, M.A. & Bushman, C., 2008. Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of theoretical biology*, 254(3), pp.541–5.

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

Jensen, K., Call, J. & Tomasello, M., 2007b. Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences of the United States of America*, 104(32), pp.13046–50.

Johansson-Stenman, O., Mahmud, M. & Martinsson, P., 2005. Does stake size matter in trust games? *Economics Letters*, 88(3), pp.365–369.

Johnson, D.D.P., Stopka, P. & Knights, S., 2003. Sociology: The puzzle of human cooperation. *Nature*, 421(6926), pp.911–2.

Johnson, T. et al., 2009. The role of egalitarian motives in altruistic punishment. *Economics Letters*, 102(3), pp.192–194.

Jordan, J.J., McAuliffe, K. & Rand, D.G., 2014. Third-Party Punishment is Motivated by Anger and is not an Artifact of Self-Focused Envy or the Strategy Method. *SSRN Electronic Journal*.

Kagel, J.H., Kim, C. & Moser, D., 1996. Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior*, 13(1), pp.100–110.

Kagel, J.H. & Roth, A.E., 1995. *The Handbook of Experimental Economics*,

Keser, C. & van Winden, F., 2000. Conditional Cooperation and Voluntary Contributions to Public Goods. *Scandinavian Journal of Economics*, 102(1), pp.23–39.

Killingback, T. & Doebeli, M., 2002. The continuous prisoner's dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *The American naturalist*, 160(4), pp.421–38.

Killingback, T., Doebeli, M. & Knowlton, N., 1999. Variable investment, the Continuous Prisoner's Dilemma, and the origin of cooperation. *Proceedings. Biological sciences / The Royal Society*, 266(1430), pp.1723–8.

Kocher, M.G., Martinsson, P. & Visser, M., 2008. Does stake size matter for cooperation and punishment? *Economics Letters*, 99(3), pp.508–511.

Kube, S. & Traxler, C., 2010. The Interaction of Legal and Social Norm Enforcement. *CESifo Working Paper Series*.

Kümmerli, R., 2011. A test of evolutionary policing theory with data from human societies. M. Perc, ed. *PloS one*, 6(9), p.e24350.

Lamba, S. & Mace, R., 2011. Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(35), pp.14426–30.

Ledyard, J.O., 1995. Public Goods: A Survey of Experimental Research. In *The Handbook of Experimental Economics*. pp. 111–194.

Lehmann, L. et al., 2007. Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. *The American naturalist*, 170(1), pp.21–36.

Lehmann, L., Feldman, M.W. & Rousset, F., 2009. On the evolution of harming and recognition in finite panmictic and infinite structured populations. *Evolution; international journal of organic evolution*, 63(11), pp.2896–913.

Leimar, O. & Hammerstein, P., 2010. Cooperation for direct fitness benefits. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1553), pp.2619–26.

Levitt, S.D. & List, J.A., 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives*, 21(2), pp.153–174.

Loomes, G., 1999. Some Lessons from Past Experiments and Some Challenges for the Future. *Economic Journal*, 109(453), pp.F35–45.

Luce, R.D. & Raïffa, H., 1957. *Games and Decisions: Introduction and Critical Survey*, Courier Dover Publications.

Mahdi, N.Q., 1986. Pukhtunwali: Ostracism and honor among the Pathan Hill tribes. *Ethology and Sociobiology*, 7(3-4), pp.295–304.

Marlowe, F.W. & Berbesque, J.C., 2008. More "altruistic" punishment in larger societies. *Proceedings. Biological sciences / The Royal Society*, 275(1634), pp.587–590.

Mason, W. & Suri, S., 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1), pp.1–23.

Mathew, S. & Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences of the United States of America*, 108(28), pp.11375–80.

Maynard Smith, J., 1982. *Evolution and the Theory of Games*, Cambridge University Press.

McCabe, K.A., Rigdon, M.L. & Smith, V.L., 2003. Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52, pp.267–275.

McCullough, M.E., Kurzban, R. & Tabak, B. a, 2013. Cognitive systems for revenge and forgiveness. *The Behavioral and brain sciences*, 36(1), pp.1–15.

McNamara, J.M. et al., 2008. The coevolution of choosiness and cooperation. *Nature*, 451(7175), pp.189–92.

Milinski, M., Semmann, D. & Krambeck, H.-J., 2002. Reputation helps solve the "tragedy of the commons". *Nature*, 415(6870), pp.424–6.

Minson, J.A. & Monin, B., 2012. Do-Gooder Derogation: Disparaging Morally Motivated Minorities to Defuse Anticipated Reproach. *Social Psychological and Personality Science*, 3, pp.200–207.

Monin, B., Sawyer, P.J. & Marquez, M.J., 2008. The rejection of moral rebels: resenting those who do the right thing. *Journal of personality and social psychology*, 95, pp.76–93.

Monroe, K.B. & Lee, A.Y., 1999. Remembering versus Knowing: Issues in Buyers' Processing of Price Information. *Journal of the Academy of Marketing Science*, 27(2), pp.207–225.

El Mouden, C. et al., 2014. Cultural transmission and the evolution of human behaviour: a general approach based on the Price equation. *Journal of Evolutionary Biology*, 27(2), pp.231–241.

Munier, B. & Zaharia, C., 2002. High Stakes and Acceptance Behavior in Ultimatum Bargaining. *Theory and Decision*, 53(3), pp.187–207.

Nash, J., 1951. Non-Cooperative Games. *Annals of Mathematics*, 54(2), pp.286–295.

Neiworth, J.J. et al., 2009. Is a sense of inequity an ancestral primate trait? Testing social inequity in cotton top tamarins (Saguinus oedipus). *Journal of comparative psychology*, 123, pp.10–17.

Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: Can we really govern ourselves. *Journal of Public Economics*, pp.91 – 112.

Nikiforakis, N. & Engelmann, D., 2008. Feuds in the Laboratory? A Social Dilemma Experiment. *Department of Economics - Working Papers Series*.

Nikiforakis, N. & Normann, H.-T., 2008a. A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11(4), pp.358 – 369.

Nikiforakis, N. & Normann, H.-T., 2008b. A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11(4), pp.358–369.

Nikiforakis, N., Normann, H.-T. & Wallace, B., 2010. Asymmetric Enforcement of Cooperation in a Social Dilemma. *Southern Economic Journal*, 76(3), pp.638–659.

Nikiforakis, N., Noussair, C.N. & Wilkening, T., 2012. Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, 96(9-10), pp.797–807.

Nikiforakis, N., Oechssler, J. & Shah, A., 2014. Hierarchy, coercion, and exploitation: An experimental analysis. *Journal of Economic Behavior & Organization*, 97(C), pp.155–168.

Noë, R. & Hammerstein, P., 1995. Biological markets. *Trends in ecology & evolution (Personal edition)*, 10, pp.336–339.

Normann, H.-T. & Wallace, B., 2012. The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *International Journal of Game Theory*, 41(3), pp.707–718.

Noussair, C. & Tan, F., 2011. Voting on Punishment Systems within a Heterogeneous Group. *Journal of Public Economic Theory*, 13(5), pp.661–693.

Nowak, M.A. & Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685), pp.573–7.

Nowak, M.A. & Sigmund, K., 2005. Evolution of indirect reciprocity. *Nature*, 437(7063), pp.1291–8.

Ostrom, E., Walker, J. & Gardner, R., 1992. Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review*, 86(2), pp.404–417.

Page, T., Putterman, L. & Unel, B., 2005. Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency. *The Economic Journal*, 115(506), pp.1032–1053.

Panchanathan, K. & Boyd, R., 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432(7016), pp.499–502.

Parks, C.D. & Stone, A.B., 2010. The desire to expel unselfish members from the group. *Journal of personality and social psychology*, 99(2), pp.303–10.

Pedersen, E.J., Kurzban, R. & McCullough, M.E., 2013. Do humans really punish altruistically? A closer look. *Proceedings. Biological sciences / The Royal Society*, 280(1758), p.20122723.

Peysakhovich, A., Nowak, M.A. & Rand, D.G., 2014. Humans display a "cooperative phenotype" that is domain general and temporally stable. *Nature communications*, 5, p.4939.

Prediger, S., Vollan, B. & Herrmann, B., 2014. Resource scarcity and antisocial behavior. *Journal of Public Economics*, 119, pp.1–9.

Prediger, S., Vollan, B. & Herrmann, B., 2013. Resource Scarcity, Spite and Cooperation. , (227).

Price, M.E., Cosmides, L. & Tooby, J., 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, 23(3), pp.203–231.

Przepiorka, W. & Diekmann, A., 2013. Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proc Biol Sci*, 280(1759), p.20130247.

De Quervain, D.J.-F. et al., 2004. The neural basis of altruistic punishment. *Science (New York, N.Y.)*, 305(5688), pp.1254–8.

R Development Core Team, R., 2011. R: A Language and Environment for Statistical Computing R. D. C. Team, ed. *R Foundation for Statistical Computing*, 1(2.11.1), p.409.

Raihani, N.J., McAuliffe, K., et al., 2012. Are cleaner fish, Labroides dimidiatus, inequity averse? *Animal Behaviour*, 84(3), pp.665–674.

Raihani, N.J., 2014. Hidden altruism in a real-world setting. *Biology letters*, 10, p.20130884.

Raihani, N.J., Pinto, A.I., et al., 2012. Male cleaner wrasses adjust punishment of female partners according to the stakes. *Proceedings. Biological sciences / The Royal Society*, 279(1727), pp.365–70.

Raihani, N.J. & Bshary, R., 2015. The reputation of punishers. *Trends in Ecology & Evolution*.

Raihani, N.J., Grutter, A. & Bshary, R., 2012. Female cleaner fish cooperate more with unfamiliar males. *Proc Biol Sci*.

Raihani, N.J., Grutter, A.S. & Bshary, R., 2010. Punishers benefit from third-party punishment in fish. *Science (New York, N.Y.)*, 327(5962), p.171.

Raihani, N.J. & Hart, T., 2010. Free-riders promote free-riding in a real-world setting. *Oikos*, 119(9), pp.1391–1393.

Raihani, N.J., Mace, R. & Lamba, S., 2013. The effect of $1, $5 and $10 stakes in an online dictator game. *PloS one*, 8(8).

Raihani, N.J. & McAuliffe, K., 2012a. Does Inequity Aversion Motivate Punishment? Cleaner Fish as a Model System. *Social Justice Research*, 25(2), pp.213–231.

Raihani, N.J. & McAuliffe, K., 2012b. Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, (July), pp.18–21.

Raihani, N.J., Thornton, A. & Bshary, R., 2012. Punishment and cooperation in nature. *Trends in ecology & evolution*, 27(5), pp.288–95.

Rand, D.G. et al., 2009. Positive interactions promote public cooperation. *Science (New York, N.Y.)*, 325(5945), pp.1272–5.

Rand, D.G. et al., 2014. Social heuristics shape intuitive cooperation. *Nature communications*, 5, p.3677.

Rand, D.G., 2012. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*, 299(null), pp.172–9.

Rand, D.G., Greene, J.D. & Nowak, M.A., 2012. Spontaneous giving and calculated greed. *Nature*, 489(7416), pp.427–30.

Rankin, D.J., Dos Santos, M. & Wedekind, C., 2009. The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50), p.E135; author reply E136.

Rapoport, A. & Chammah, A.M., 1965. *Prisoner's dilemma: a study in conflict and cooperation*, Ann Arbor, Univ. of Michigan Press.

Rapoport, A. & Dale, P.S., 1966. The "end" and "start" effects in iterated Prisoner's Dilemma. *Journal of Conflict Resolution*, 10(3), pp.363–366.

Reuben, E. & Riedl, A., 2013. Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1), pp.122–137.

Ridley, M., 2004. *Evolution*,

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

Roberts, G., 2013. When punishment pays. *PloS one*, 8(3), p.e57378.

Roberts, G. & Sherratt, T.N., 1998. Development of cooperative relationships through increasing investment. *Nature*, 394(6689), pp.175–9.

Rockenbach, B. & Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444(7120), pp.718–723.

Roos, P. et al., 2014. High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings. Biological sciences / The Royal Society*, 281(1776).

Ross, J. et al., 2010. Who are the Crowdworkers?: Shifting Demographics in Amazon Mechanical Turk. *Proc Extended Abstracts of the SIGCHI conference on Human factors in computing systems CHI*, pp.2863–2872.

Ross, L. & Ward, A., 1996. Naive realism in everyday life: Implications for social conflict and misunderstanding. Values and knowledge. In *Values and knowledge, The Jean Piaget symposium series*. pp. 103–135.

Rustagi, D., Engel, S. & Kosfeld, M., 2010. Conditional cooperation and costly monitoring explain success in forest commons management. *Science (New York, N.Y.)*, 330, pp.961–965.

Sahlins, M., 1972. *Stone Age Economics*, Aldine-Atherton.

Sanfey, A.G. et al., 2003. The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626), pp.1755–1758.

Dos Santos, M., Rankin, D.J. & Wedekind, C., 2013. Human cooperation based on punishment reputation. *Evolution*, 67(8), pp.2446–2450.

Dos Santos, M., Rankin, D.J. & Wedekind, C., 2011. The evolution of punishment through reputation. *Proceedings. Biological sciences / The Royal Society*, 278(1704), pp.371–377.

Schielzeth, H., 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2), pp.103–113.

Schultz, P.W. et al., 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological science : a journal of the American Psychological Society / APS*, 18(5), pp.429–434.

Sefton, M., Shupp, R. & Walker, J.M., 2007. The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45(4), pp.671–690.

Seinen, I. & Schram, A., 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review*, 50, pp.581–602.

Sherratt, T.N. & Roberts, G., 1998. The evolution of generosity and choosiness in cooperative exchanges. *Journal of Theoretical Biology*, 193(1), pp.167–177.

Sigmund, K., 2007. Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution*, 22(11), pp.593–600.

Sigmund, K. et al., 2011. *Social control and the social contract: The emergence of sanctioning systems for collective action*,

Sigmund, K. et al., 2010. Social learning promotes institutions for governing the commons. *Nature*, 466(7308), pp.861–3.

Sigmund, K., Hauert, C. & Nowak, M.A., 2001. Reward and punishment. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), pp.10757–62.

Silberberg, A. et al., 2009. Does inequity aversion depend on a frustration effect? A test with capuchin monkeys (Cebus apella). *Animal Cognition*, 12, pp.505–509.

Slonim, R. & Roth, A.E., 1998. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. *Econometrica*, 66, pp.569–596.

Smith, P. & Silberberg, A., 2010. Rational maximizing by humans (Homo sapiens) in an ultimatum game. *Animal cognition*, 13(4), pp.671–7.

Steiner, J., 2007. A Trace of Anger is Enough: On the Enforcement of Social Norms. *CERGE-EI Working Papers*.

Strobel, A. et al., 2011. Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage*, 54(1), pp.671–80.

Suri, S. & Watts, D.J., 2011. Cooperation and contagion in web-based, networked public goods experiments. *PloS one*, 6(3).

Sutter, M., Kocher, M. & Strauß, S., 2003. Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, 81(3), pp.341–347.

Sylwester, K., Herrmann, B. & Bryson, J., 2013. Homo homini lupus? Explaining antisocial punishment. *British Educational Research Journal*, 6, pp.167–185.

Sylwester, K. & Roberts, G., 2010. Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, 6(5), pp.659–662.

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

Symonds, M.R.E. & Moussalli, A., 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65, pp.13–21.

Tan, F., 2008. Punishment in a Linear Public Good Game with Productivity Heterogeneity. *De Economist*, 156(3), pp.269–293.

Thomas, M. & Morwitz, V., 2005. Penny Wise and Pound Foolish: The Left-Digit Effect in Price Cognition. *Journal of Consumer Research*, 32(1), pp.54–64.

Tooby, J., Cosmides, L. & Price, M.E., 2006. Cognitive Adaptations for n-person Exchange: The Evolutionary Roots of Organizational Behavior. *Managerial and decision economics : MDE*, 27(2-3), pp.103–129.

Traulsen, A., Röhl, T. & Milinski, M., 2012. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings. Biological sciences / The Royal Society*, 279(1743), pp.3716–21.

Trivers, R.L., 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), pp.35–57.

Úbeda, F. & Duéñez-Guzmán, E.A., 2011. Power and corruption. *Evolution; international journal of organic evolution*, 65(4), pp.1127–39.

Vukov, J. et al., 2013. Reward from punishment does not emerge at all costs. S. A. Levin, ed. *PLoS computational biology*, 9(1).

Vyrastekova, J., Funaki, Y. & Takeuchi, A., 2008. Strategic vs. Non-Strategic Motivations of Sanctioning. *SSRN Electronic Journal*.

Wahl, L.M. & Nowak, M.A., 1999. The continuous Prisoner's dilemma: II. Linear reactive strategies with noise. *Journal of Theoretical Biology*, 200(3), pp.307–321.

Waite, T.A. & Campbell, L.G., 2006. Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience*, 13(4), pp.439–442.

Walker, J.M. & Halloran, M.A., 2004. Rewards and Sanctions and the Provision of Public Goods in One-Shot Settings. *Experimental Economics*, 7(3), pp.235–247.

Wang, C.S. et al., 2011. Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organizational Behavior and Human Decision Processes*, 116, pp.46–54.

Wang, R. et al., 2010. Asymmetric interaction will facilitate the evolution of cooperation. *Science China. Life sciences*, 53(8), pp.1041–6.

Wedekind, C. & Braithwaite, V.A., 2002. The long-term benefits of human generosity in indirect reciprocity. *Current Biology*, 12(12), pp.1012–5.

Wellington, H., 1976. *Union Fines and Workers' Rights*,

West, S., Griffin, A. & Gardner, A., 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of evolutionary biology*, 20(2), pp.415–32.

West, S., El Mouden, C. & Gardner, A., 2010. Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, 32(4), pp.231–262.

Wiessner, P., 2005. Norm enforcement among the Ju/'hoansi Bushmen. *Human Nature*, 16(2), pp.115–145.

Wu, J.-J. et al., 2009. Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41), pp.17448–51.

Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, 77(1), pp.321–344.

Xiao, E. & Houser, D., 2005. Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20), pp.7398–7401.

Yamagishi, T., 1988. Seriousness of Social Dilemmas and the Provision of a Sanctioning System. *Social Psychology Quarterly*, 51, p.32.

Yamagishi, T., 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51(1), pp.110–116.

Yu, R., Calder, A.J. & Mobbs, D., 2014. Overlapping and distinct representations of advantageous and disadvantageous inequality. *Human Brain Mapping*, 35(7), pp.3290–3301.

Zhang, B. et al., 2013. The evolution of sanctioning institutions: an experimental approach to the social contract. *Experimental Economics*.