DOCTOR OF PHILOSOPHY

The effect of social rewards and punishments on learning and cooperative decision-making

Beston, Pippa

Award date:
2019

Awarding institution:
Bangor University

PRIFYSGOL
BANGOR
UNIVERSITY

The effect of social rewards and punishments on learning and cooperative decision-making

Philippa J Beston

Thesis is submitted to the School of Psychology, Bangor University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Bangor, North Wales, United Kingdom

November 2019

Declaration

Yr wyf drwy hyn yn datgan mai canlyniad fy ymchwil fy hun yw'r thesis hwn, ac eithrio lle nodir yn wahanol. Caiff ffynonellau eraill eu cydnabod gan droednodiadau yn rhoi cyfeiriadau eglur. Nid yw sylwedd y gwaith hwn wedi cael ei dderbyn o'r blaen ar gyfer unrhyw radd, ac nid yw'n cael ei gyflwyno ar yr un pryd mewn ymgeisiaeth am unrhyw radd oni bai ei fod, fel y cytunwyd gan y Brifysgol, am gymwysterau deuol cymeradwy.

I hereby declare that this thesis is the results of my own investigations, except where otherwise stated. All other sources are acknowledged by bibliographic references. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless, as agreed by the University, for approved dual awards.

Acknowledgements

Firstly, I would like to thank my supervisors for steering me through the increasingly challenging waters of my PhD. To Dr Erin Heerey, thank you so much for giving me this opportunity; I have no idea what you saw in that oddball who wandered into your lab all those years ago. But thank you for investing so much time and energy into me, for the truthful feedback, and for teaching me all of the skills that I now take for granted. Being 5 time zones away hasn't been easy at times, but thank you so much for sticking it out with me! To Professor Guillaume Thierry, I am so very grateful for you graciously adopting me into your lab at such a late stage in my PhD. Thank you for teaching me about neuroscience, working with me through all the stages of my ERP project, and for your valuable input on other areas of my PhD. I've learned a lot from being in the BULET lab, so thank you for the opportunity to get involved. I would also like to thank the chair of my PhD committee; Professor Emily Cross – you have been wonderfully supportive and helpful throughout this process.

I would also like to thank other members of the BULET lab for their immense generosity with their time: Dr Cécile Barbet and Yang Li. Thanks should also be extended to previous members of the Bangor Heerey Lab: Dr Thandi Gilder and Dr Danni Shore, who were such great role models when I was just starting to think about graduate school. Special thanks go to Thandi, for being such an awesome person to work with, and for your kind and patient ear.

I would also like to thank my marvellous friends, I don't know where I'd be without you all. Thank you to Jez for the long discussions about the patriarchy, large wine glasses, and for making me a better feminist. Joe, for accompanying me on my brief jaunt into student politics. To Helen, Claudia and Ciaran, for the evenings of fun, board games and beige food. Polly, for listening to my seemingly unending stream of woes over dinner; for the science chats and all of the PhTea. Louise (PhD wife), for the working-weekends, listening to my rants, dealing with my multitude of crises, and for feeding me inordinate amounts of cheese. Lisa, for the random Bangor trips, giggles, and the financial advice. To Alyssa, for the long-distance sanity check-ups and soothing purrs. To Ben, for always being there in the background; watching and listening. And finally, to my best friend Adele, for the long conversations about how everything will be ok. You have all really helped me through this, even if you weren't aware of it at the time. My gratitude to you, always.

Finally, I would like to thank my parents for all of their support – support that has taken many forms over the years (especially financial!). Thank you for encouraging me to go to university in the first place and for standing by all of the decisions that I've made since I've been at Bangor.

Table of Contents

List of Figures

List of Tables

Thesis Summary

This thesis explores how exposure to social information affects decision-making in two different domains. In one, I use a unique variant of the Public Goods Game (PGG) paradigm to examine how social and monetary punishments and rewards alter decisions to cooperate in the interpersonal domain. At the intrapersonal level, I examine how exposure to social rewards affects implicit learning (and the neural signature thereof) in a novel task.

In the first two empirical chapters, I compared how the opportunity to interact face-to-face, versus traditional anonymous interactions, influenced cooperative decision-making in the PGG. This work additionally examined whether different punishment types (i.e., social-reputational and/or monetary) affected contribution behaviour across interaction context. In the anonymous context, monetary sanctions were the most effective in promoting cooperation. However, in the face-to-face context, social punishments were most effective. Additionally, group interaction positivity on a given trial predicted investment on the next.

In the second empirical chapter, I examined the effects of rewards on cooperation. In contrast to punishment, monetary rewards were more effective in maintaining contributions than social rewards in the face-to-face context. In this case, the incentive to gain a good reputation may have been so explicit that it 'crowded-out' cooperation.

In the final empirical chapter, I asked participants to implicitly learn the rules of a novel card game and compared the effectiveness of social versus non-social feedback. Here, I used event-related potential (ERP) methodology to examine the neural markers of learning. After learning, we found differences in the strength of ERP components between the social and non-social feedback groups across task conditions. These results suggest that socially salient feedback alters the process of implicit learning.

Together, this work shows that exposure to social information affects cooperative decision-making during the Public Goods Game, and also alters the neural signature of implicit learning. Thus, although challenging to capture, this thesis has begun to account for the social factors that affect people's every-day decisions.

Chapter 1

General Introduction

Thesis overview

Decisions about how to behave in the presence of others and how to act in social contexts are made on a moment-to-moment basis every day. We frequently face the choice of whether to take more than our fair share of the pie; to work collaboratively with co-workers; or to snub those who have treated us unfairly in the past. Often, these decisions are made in the context of face-to-face social interactions and they are likely influenced by a wealth of social cues. This social information often affects decision-making; influencing the allocation of attentional resources, evaluations of others and even specific decision results (Shore & Heerey, 2013). Furthermore, social information is highly relevant to decision-making and has ostensibly shaped human cognitive processes (Cosmides, 1989; Cosmides & Tooby, 1995). Indeed, humans have likely evolved to be good social partners, and so behaviour often reflects choices that signal this (Everett, Pizarro, & Crockett, 2016). Additionally, being a good social partner also often means cooperating with others in the social world.

Cooperation is the cornerstone of a smooth-running society. We require neighbours and co-workers to cooperate with each other, politicians to cooperate with their constituents, and heads of government to cooperate with their international counterparts. At the heart of these situations is a tension between group welfare and an individual's personal benefit. Individuals are often better off when they prioritise their own bottom line, but this is often to the detriment of those around them. On the other hand, a rising tide benefits everyone; group welfare is often collectively better when everyone contributes their fair share (Rand, Dreber, Ellingsen, Fundenberg, & Nowak, 2009). For example, provisions to public services (i.e., national defence, healthcare) is much better off when everyone pays the taxes that they owe, even if an individual can still access these services if they choose not to pay their fair share.

Thus, the motivations behind cooperation and defection are both interesting to study and have important real-world implications.

One method of studying social dilemmas within a laboratory context is via economic games. Here, the tension between maximising personal profit and the collective interest is captured within elegant incentive structures. These games thus represent a microcosm of human behaviour and decision-making. However, predictions about how people should behave during these games, via economic models, are often naïve when compared to empirical observations of actual human behaviour (Camerer & Fehr, 2006). People are often more cooperative than expected within laboratory contexts, especially when interactions are repeated, or reputation is at stake (Rand & Nowak, 2013). Additionally, the inclusion of certain mechanisms, such as monetary rewards and punishments, further incentivise cooperation (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). However, as yet, little is known about how different *types* of rewards and punishments may differentially affect cooperation, depending on the environment in which they are administered. Furthermore, experiments in this area typically involve anonymised interactions between group members (e.g., Fehr & Gächter, 2002); where people are unable to see or interact with their social partners face-to-face. Thus, we do not know the effect that naturalistic social interactions may have on the propensity to cooperate with others, especially if these interactions are positive. As a result, we explore how the efficacy of different types of rewards and punishments changes depending on the social environment in empirical Chapters 2 and 3.

Additionally, it is important to note that social feedback from interaction partners is prevalent in the environment, as well as being important in altering our social behaviours. Furthermore, social rewards (i.e., smiles) are subjectively more valuable than non-social alternatives, even when both stimulus types constitute the same objective value (Heerey, 2014; Shore & Heerey, 2011). Furthermore, when used as feedback, social stimuli appear to

be more effective than non-social feedback during associative learning tasks (Hurlemann et al., 2010). However, the efficacy of social feedback during implicit learning tasks is thus so far unknown. Additionally, research has suggested that the Event-Related Potential (ERP) technique is an effective method for examining the process of implicit learning (Baldwin & Kutas, 1997), however, it is also unclear whether the type of feedback provided during this task may affect the neural signature of learning. Thus, in the last chapter of this thesis (Chapter 4), we explore whether social feedback, relative to non-social feedback, differentially affects behavioural performance or the neural signature of learning, during a novel implicit learning task.

Thus, in this thesis, we take advantage of the diversity of methods used in social decision-making research to examine the effects of socially-relevant feedback during cooperative decision-making and implicit learning.

Game theory and human behaviour

In models of economic decision-making, classic theories of behaviour have suggested that humans should make 'rational' decisions that maximise utility for themselves (von Neumann & Morgenstern, 1944)[1]. Rational individuals are assumed to have exclusively self-regarding preferences (i.e. those that do not account for the outcomes of others); to have correct and complete information about their situation and opponents; and to be able to perfectly compute the best course of action to satisfy their self-interest (Camerer & Fehr, 2006). This idea outcomes is sometimes referred to as 'homo-economicus' or 'economic man' (Hollander, 2000).

[1] It is worth noting here that throughout this document, when I use the term 'rationality' I am referring to this definition, as opposed to other definitions of rationality, for instance 'ecological rationality', which means adopting a strategy that best fits the context in which an actor is making a decision (Goldstein & Gigerenzer, 2002)

Accordingly, humans as decision makers are conceptualised as selfishly inclined (Dewall, Baumeister, Gailliot, & Maner, 2008). However, in laboratory settings, they are demonstrably less rational than economic theories might predict (Camerer & Fehr, 2006; Henrich et al., 2001). This may be due to limitations in human's cognitive resources, meaning that their decision-making is subject to 'bounded rationality' – limitations to outcome computations arising from limited cognitive and working memory capacity (Camerer, 1998; Landa & Wang, 2001). In practice, then, people are much more cooperative than many theoreticians suggest that they should be. Thus, cooperation, whilst sacrificing economic payoff, is a highly replicable phenomenon observed in many psychological/ experimental economics reports (Andreoni, 1995; Andreoni & Miller, 1993; Fehr & Gächter, 2000b; Marwell & Ames, 1981; Ostrom, 2000).

Thus, humans are often assumed to be able to determine the most self-regarding outcomes and behave in self-maximising ways. However, the above examples suggest that this is not always the case. Decisions taken in the social world are rarely without implications for others. Indeed, people learn about the world and its give-and-take contingencies from their social interaction partners; gaining feedback from others that shapes their own behaviour (Heerey & Velani, 2010). These learned behavioural standards then appear to become embedded in social behaviours (Chudek & Henrich, 2011; Peysakhovich & Rand, 2016), and become intuitive 'rules of thumb' that arise from existing and interacting with others in human society (see the 'Social Heuristics Hypothesis': Rand, 2016; Rand, Peysakhovich, et al., 2014)[2]. Unsurprisingly, therefore, intuitionist models of decision-

[2] However, this hypothesis and evidence has been disputed by Tinghög et al., (2013), after failing to replicate some of the original findings that formed the basis of this hypothesis in Rand, Greene, & Nowak (2012). Tinghög et al., instead suggested that the original findings were an artefact of the exclusion of participants in the original studies, and found in their own replication attempts that time pressure did not lead to increased cooperation, as hypothesised by Rand et al. However, in a reply to these critiques, Rand et al., re-analysed this dataset without excluding subjects in the same manner and continued to find their expected effect. However,

making appear to better explain human behaviour than rationality (Haidt, 2001). The idea of humans being less than rational is highlighted in the disparity between game-theoretical predictions of behaviour and actual behaviour in economic games.

Nonetheless, formal game theory began by looking at dilemmas between dyads, such as the Prisoner's Dilemma, which was effective at modelling 'rational' decision-making between pairs of people (Myerson, 1991). These models helped to highlight theoretically 'optimal' behaviour, even if they were not necessarily meant to reflect real life decision-making processes (Poundstone, 1992). However, from the paradigm of the Prisoner's Dilemma, some limitations arose, including that it only attempts to model interactions between two agents. In real life, social interactions are not always dyadic - people often interact in groups, involving a dynamic network of agents (Rand, Arbesman, & Christakis, 2011).

Why study economic games? Oftentimes, individuals must decide whether to favour their own interests over that of the collective. These decisions occur on many different scales ranging from, for example, individual decisions to recycle, or to refrain from using a hosepipe during drought; to agreements across nations to take action against climate change (e.g., the Paris Agreement) or whether an individual nation should continue to cooperate with economic and political structures that bring peace across a continent (e.g., the EU). These so called 'social dilemmas' (Dawes, 1980; Van Lange & Joireman, 2008) can be boiled down and studied in the lab in order to provide answers about how to foster cooperation in real life situations (Kraft-Todd, Yoeli, Bhanot, & Rand, 2015), and provide scalable solutions for global political problems (e.g., Rand, Yoeli, & Hoffman, 2014). This therefore makes the study of social dilemmas an important and highly relevant topic for social scientists.

---

given that Tinghög did fail to replicate the original studies suggests that we should exercise caution when considering the 'intuitive cooperation' idea.

The realism of economic games. However, it also important to take into consideration some limitations to how these social dilemmas are studied in the laboratory. Indeed, the realism of experimental economic games have been under dispute since they have been implemented in psychology (Camerer, 2011). Those who have been critical of their usage have cited several problems in the generalisability of findings from economic games beyond the lab context in which they are typically played. For instance, there is the potential for self-selection bias in the samples recruited for University lab experiments, leading to limited applications outside of the this particular context; especially as demographics of these samples may be different to those participating in real market environments (Levitt & List, 2007). Additionally, players may bring to the lab their own pre-play experiences shaped by real life – a factor that is difficult to control a-priori (Levitt & List, 2007). Experimental research corroborates that participants do indeed play according to their own heuristics, but also that this is a factor that can be manipulated and studied in the laboratory environment (Peysakhovich & Rand, 2016; Rand, Peysakhovich, et al., 2014). Levitt and List (2007) further suggest that the timeframe in which players are given to make decisions in the laboratory is much shorter than that of real life, therefore not accurately representing real world decision-making processes. Subsequent research has indeed found that contribution decisions made over a longer time frame are typically less cooperative (Rand, Greene, & Nowak, 2012).

Camerer (2011) countered these criticisms by arguing that the field of experimental economics was designed to link economic theory to behaviour, and so generalisability was never the goal of these experiments. Levitt and List (2007) also concede that it is useful to test a theory in a 'clean' and controlled environment, such as the laboratory. Furthermore, (Camerer, 2011) also argued that in some cases, lab findings *do* generalise to field studies - for example, one study found that cheating in a laboratory setting predicted cheating

behaviour in the field (Potters & Stoop, 2016). Thus, typical lab features do not necessarily undermine generalisability, as there appears to be a link between participant behaviour in both laboratory and field settings. Therefore, it is likely that lab experiments can contribute to our *general* understanding of human behaviour (Camerer, 2011), even if it is not under the exact same conditions as the real world.

The Public Goods Game. One such social dilemma of particular interest to social and political scientists, and economists is the Public Goods dilemma. Public goods situations rely on the provisions of others to exist; individual contributions are optional and these resources are then evenly distributed across individuals (Gravelle & Rees, 2004). These provisions add value to others' lives and ensure that goods and services are available to the many, examples including: public broadcasting services, software development and, nationalised health-care services (e.g., the NHS in the UK). Public goods situations also arise in everyday settings, such as having a clean and tidy living space in shared accommodation and, in science, agreeing to peer-review manuscripts submitted to journals.

Public goods are characterised by being 'non-rivalrous', meaning that the consumption of the good by one individual does not reduce the amount of good that is available to another (Gravelle & Rees, 2004), for example, one individual accessing a radio broadcast does not prevent another from doing so. Similarly, they are 'non-excludable', meaning that these provisions are accessible to everyone, regardless of an individual's original contribution to that particular resource. This means that even those who have not, or cannot, contribute a "fair" share, can benefit from the provisions of others. Thus, this presents tension between private and public interests, as individuals stand to gain more by not contributing to these resources, and yet these resources fail without individual contributions of time, energy or money. This is known as the 'free-rider' problem (Baumol, 2004), which is

characterised by high contributors subsidising the benefits of those who contribute little, or none, of their own resources (Marwell & Ames, 1980).

The Public Goods Game is an effective way to examine human cooperation and the propensity of free-riding behaviour in a laboratory context. In these experimental games, participants receive an endowment of monetary tokens and must choose a contribution amount to a public fund. The fund is then tallied, multiplied, and split evenly amongst the players, regardless of initial contributions (Andreoni, 1995).

According to economic models, a contribution of zero is the dominant strategy in the public goods game (Olson, 1965, as cited in Ostrom, 2000). However, evidence suggests that players are often more cooperative than "rational" models would predict (Camerer & Fehr, 2006; Kahneman & Tversky, 1979; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991), showing that players are willing to sacrifice their own personal payoffs in favour of the group's benefit (Isaac & Walker, 1988; Isaac et al., 1994; Rand, Dreber, Ellingsen, Fundenberg, & Nowak, 2009). Research has suggested that players may be intuitively cooperative, based on their previous social interactions where cooperation has been beneficial (Rand et al., 2014; Rand, Greene, & Nowak, 2012), and thus, cooperation may be a default state due to repeated interactions with social group members, (Andreoni, 1988). However, after experience in the laboratory context, where cooperation may not be necessarily as beneficial between strangers, cooperation declines (Andreoni & Croson, 2008; Burton-Chellew & West, 2013; Burton-Chellew, Nax, & West, 2015; Croson, 1996; Fehr & Gächter, 2000). Moreover, in laboratory settings, there are often negligible real-world repercussions following defection, further incentivising free-riding behaviour (Guala, 2012).

As a result, laboratory studies have been interested in the question of how cooperation can be maintained in the Public Goods Game. In this game, it is not possible to reciprocate against free-riders directly, in the same way that players can engage in tit-for-tat strategies in

the dyadic Prisoner's Dilemma game (Axelrod & Hamilton, 1981). It is of course possible for a player to reduce their own contribution to the public good as a response to another free-riding player lowering his/her contribution, however, this hurts the welfare of the group and eventually leads to mutual defection (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). Therefore, laboratory studies have examined the introduction of mechanisms designed to encourage cooperation. These mechanisms include costly punishments and costly rewards, and are often (but not exclusively) targeted at non-cooperative (or 'free-riding') players and cooperative players, respectively.

Incentives to maintain cooperation

Costly punishment. In order to maintain contributions to the public good, experimental games have studied several mechanisms, one such being costly punishment (Fehr & Gächter, 2000a, 2002). Costly punishment is the act of incurring a cost to oneself in order to incur a cost to another, typically, a free-riding player, (costly punishment is also known as 'altruistic punishment'). Costly punishment is readily employed by participants (Nikiforakis & Normann, 2008). Indeed, people even prefer institutions where it is possible to sanction others as opposed to those where it is not (Gürerk, Irlenbusch, & Rockenbach, 2006). Group contributions are typically larger under the threat of sanctions than when this threat is not present (Chaudhuri, 2010; Sefton, Shupp, & Walker, 2007). However, to maintain contributions over time, it is also important that the punishment is cost-effective – i.e. that the cost to implement punishment is sufficiently low and the impact on the recipient's payoff sufficiently high (Chaudhuri, 2010; Egas & Riedl, 2008; Nikiforakis & Normann, 2008).

Although costly punishment is readily used by participants[3], and people even opt to pay additional costs to conceal punishment (Rockenbach & Milinski, 2011), it is not an economically 'rational' option. This behaviour requires an individual to bear the cost of punishment when the benefits are often unclear (Fehr & Gächter, 2002; Egas & Riedl, 2008). Thus, this behaviour is often regarded as puzzling, especially when it is observed in situations in which punishers may not reap the rewards of reforming defectors down the road (i.e., in 'one-shot' games; Rand, Greene, & Nowak, 2012; Walker & Halloran, 2004). Perhaps this propensity for punishment is a form of emotional self-expression (Xiao & Houser, 2005) or comes from a desire for retribution (Crockett, Özdemir, & Fehr, 2014). Thus, researchers have explained this phenomenon in terms of the utility players gain, in the form of pleasure or satisfaction from the act of punishing (de Quervain et al., 2004; Singer et al., 2006)[4].

Interestingly, there is also a risk that angry free-riders receiving punishment choose to punish in retaliation, resulting in a reduction of payoffs for all group members (Hopfensitz & Reuben, 2009). And so, whilst costly punishment is effective, it can also be destructive. If used antisocially (that is, free-riders punishing co-operators), punishments can backfire, serving to reduce levels of cooperation (Herrmann, Thöni, & Gächter, 2008), smothering the potential for reciprocity between partners (Fehr & Rockenbach, 2003), and lowering payoffs for the whole group (Dreber, Rand, Fudenberg, & Nowak, 2008; Egas & Riedl, 2008; Szolnoki & Perc, 2010).

---

[3] Costly punishment behaviour also occurs in several different social contexts in animal societies. For example, European moorhens may inflict damage or even kill their young for persistently asking for food, to discourage its siblings from being greedy. This kind of parent-offspring conflict is one of many examples where costly punishment behaviour may arise in animals, as described by Clutton-Brock & Parker (1995).

[4] It is important to note here that emotional expressivity or 'spite' are proximate explanations for why costly punishment behaviour may arise, that being - an explanation as to *how* a behaviour works (Scott-Phillips, Dickins, & West, 2011). These examples cannot explain *why* the behaviour exists in the first place – which would be an ultimate explanation for the behaviour.

Costly reward. The costly rewarding of cooperators, on the other hand, has been shown to reduce the detrimental retaliatory behaviour that punishment can provoke (Dreber et al., 2008; Nikiforakis, 2008). Moreover, group earnings are often healthier when players can reward cooperative behaviour than punish free-riding behaviour (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). As with punishment, however, costly rewarding is not a rational behaviour due to the costs associated with dispensing incentives and the unknown benefits associated with doing so (Andreoni, Harbaugh, & Vesterlund, 2003). Regardless, players are indeed prepared to reward others for cooperative behaviour (Almenberg, Dreber, Apicella, & Rand, 2011) and may even prefer using rewards to punishments (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Molenmaker, De Kwaadsteniet, & Van Dijk, 2014; Sutter, Haigner, & Kocher, 2010).

Costly rewarding has the disadvantage, however, that it does not target or alter the behaviour of non-cooperative players (Andreoni et al., 2003). This means that rewards require the continued sacrifice of resources to ensure cooperation. On the other hand, once free-riding has been eliminated via punishment, players no longer need to pay to dispense these sanctions (Szolnoki & Perc, 2010). However, in terms of overall cooperation, rewards do appear to have a positive effect compared to when this incentive is not available (Balliet, Mulder, & Van Lange, 2011).

Repeated interactions

Walker & Halloran (2004) found that during one-shot interactions, rewards and punishments were not successful in raising contributions in a Public Goods-like environment. This led the authors to suggest that one influential factor in maintaining cooperation is repeated interactions with the same group members; otherwise the threat of punishment or the

promise of reward was effectively empty over the long-term [5]. Indeed, research also suggests that cooperation increases when dilemmas are iterated with consistent group members (Balliet et al., 2011). The opportunity for 'targeted interactions'; the ability to punish or reward specific group members based on their contribution amounts, are therefore a key factor that can influence cooperation, as they introduce consequences for current behaviour (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009).

Repeated interactions may also help to explain why players are willing to bear costs to punish free-riders or reward cooperative players. People often live 'in the shadow of the future' (Hauser, Rand, Peysakhovich, & Nowak, 2014; Rand, Dreber, Ellingsen, Fundenberg, et al., 2009; Rand & Nowak, 2013), meaning that interactions are often repeated in the real world, with consequences occurring down the line. In the laboratory, games that involve repeated interactions with the same partners simulate the opportunities available in real world interactions (Rand & Nowak, 2013).

In repeated interactions with consistent group members, the readiness to engage with costly punishment may serve as an important social signal (Raihani & Bshary, 2015). For example, it may be used to signal to other selfish players to reform their behaviour (Fehr & Fischbacher, 2004a). It may also be used to signal to other potential interaction partners that an individual is a trustworthy target for cooperation (Jordan, Hoffman, Bloom, & Rand, 2016). Moreover, costly punishment may not be as costly as it appears, as those who gain a reputation for punishment are often compensated for these costs by the cooperation from other group members (dos Santos, Rankin, & Wedekind, 2013). Therefore, it is important to note here that the ability to build a reputation is crucial to the emergence of cooperative

[5] However, it is also important to note that studies of "stranger" designs (i.e., multiple one-shot games, each with a different partner) are not completely unsuccessful at raising contributions, as long as punishment opportunities are available (Fehr & Gächter, 2000a).

behaviour (Engelmann & Fischbacher, 2009). Furthermore, punishments appear to be more effective when costly, as it seems to signal a commitment to fostering cooperation between group members (Balliet et al., 2011).

Similarly, in terms of rewarding behaviour, repeated interactions offer the opportunity for individuals to gain a reputation as a rewarder, which may be beneficial for an individual as it may signal that they are a trustworthy partner for cooperation (Almenberg et al., 2011). Indeed, having a good reputation is often rewarded by cooperation from others within and outside of a player's social groups (Semmann, Krambeck, & Milinski, 2005).

Reputation incentives also help to explain cooperation

During repeated interactions, reputation is almost always at stake when considering one's actions. Cooperation in situations where players are playing with consistent group members can be explained by the concept of 'direct reciprocity' (Axelrod & Hamilton, 1981). This is the idea that an individual should pay the cost of cooperation today to earn the trust of their partner for tomorrow; which is beneficial when another interaction is sufficiently probable (Axelrod, 1984). Cooperation is observed to increase when it is common knowledge that interactions will be repeated (Dal Bó, 2005; Dreber et al., 2008; Duffy & Ochs, 2009), and to drop off when it is apparent that interactions are coming to an end, as there is no longer a need to maintain a reputation for cooperation (Isaac & Walker, 1988; Isaac, Walker, & Thomas, 1984). Although most interactions in this case are dyadic, in real life our interactions are superimposed on a network of other potential partners who may or may not observe our cooperative behaviours. In this case, reputation building, via direct reciprocity only is not a sufficiently realistic explanation for why cooperation exists in situations where individuals may not interact with the same partner again.

Accordingly, the concept of 'indirect reciprocity' was posed to explain this phenomenon and is the idea that if A cooperates with B, then C will then cooperate with A

(Alexander, 1987). Here, cooperation is still costly, but may help that individual to acquire a reputation for being a trustworthy partner, thereby increasing the chances of others cooperating with that individual in the future (Milinski, Semmann, & Krambeck, 2002a; Rand & Nowak, 2013). Theoretically, the concept of indirect reciprocity allows cooperation to exist but only if it is possible to know or to estimate the reputation of another (i.e. their 'image score'; Nowak & Sigmund, 1998). There is experimental evidence to suggest people use the reputation information of others from their previous helping decisions, to inform their decision to cooperate with that partner, i.e., their cooperation is *conditional* upon their potential/actual partner's past behaviour (Fischbacher, Gächter, & Fehr, 2001; Seinen & Schram, 2006; Wedekind & Braithwaite, 2002; Wedekind & Milinski, 2000). Players with a good reputation are rewarded with cooperation from third parties who have observed their behaviour (Semmann et al., 2005).

However, researchers highlighted issues with using indirect reciprocity to explain reputation building. Unconditional cooperation remains unexplained by indirect reciprocity as this theory suggests that cooperation will only ever be conditional on whether a partner is also cooperative (Sylwester & Roberts, 2010). The concept of reputation-based partner choice was posed as an alternate explanation for the maintenance of group cooperation. This is the idea that players may develop a cooperative reputation in order to be chosen a partner for profitable interactions. This is considered to be a two-stage model of cooperative intentions, as benefits may not immediately arise in the first stage, but instead occur further down the line, i.e. at a second stage (Roberts, 1998). Knowing that another player may choose them as an interaction partner in the future increases an individual's contribution to the public good (Barclay, 2004). Additionally, the ability to *choose* a partner leads to increased contributions compared to when partners are randomly assigned (Barclay & Willer, 2007; Sylwester & Roberts, 2013). Reputation-based partner choice also increases payoff

from the public good. Players tend to prefer the most cooperative players (Barclay & Willer, 2007; Rockenbach & Milinski, 2011), and having a reputation as such pays off; high contributors gain increased access to the most profitable partnerships, thereby earning more than less cooperative players (Sylwester & Roberts, 2010). Thus, the concept of reputation-based partner choice can explain why individuals may invest in a cooperative reputation; as it is necessary to compete with others to be generous when it is possible to form cooperative (and profitable) partnerships.

In the opposite direction, research has also suggested that whilst anonymity breeds self-interested, or 'rational', behaviour (Hoffman, McCabe, Shachat, & Smith, 1994); people are more cooperative when behaviour is observable (Kraft-Todd et al., 2015; Rand, Yoeli, et al., 2014). For example, helping behaviour is observed to significantly increase when players can gain a reputation for doing so (Engelmann & Fischbacher, 2009), or when player identity is visible to other group members (Andreoni & Petrie, 2004). This effect is also apparent even when individuals *feel* that they are being watched, e.g., when stylised or real images of eyes are present during decision-making (Bateson, Nettle, & Roberts, 2006; Burnham & Hare, 2007; Ernest-Jones, Nettle, & Bateson, 2011; Haley & Fessler, 2005)[6].

Not only is this 'observability effect' present in the lab but has also been replicated in field studies. For example, more people registered for a blackout prevention program when the sign-up sheet was publicly visible in a communal area of a students' halls of residence (Yoeli, Hoffman, Rand, & Nowak, 2013). Similarly, the frequency of blood donations increases when the identity of generous donors are announced publicly (Lacetera & Macis, 2010). However, it is important to also note here that reputation mechanisms can encourage

[6] Although, it is also worth noting that this eye watching effect was not found when using a diverse subject pool, and in truly anonymous environments, such as Amazon Mechanical Turk (Raihani & Bshary, 2012). This finding was also confirmed by meta-analysis, suggesting that the effect of artificial surveillance cues on cooperation is not robustly replicable (Northover, Pedersen, Cohen, & Andrews, 2015).

cooperation, but only when cooperation is perceived positively, which can depend on the social norms of a particular group. For example, interventions using environmental sustainability incentives to 'nudge' politically conservative people to reduce electricity use appear to be less effective than when targeted at the politically liberal (Costa & Kahn, 2013). In addition, research has also suggested that although people are typically more generous in public settings, it is important that this is in the absence of monetary/material incentives – i.e., people want to be perceived as generous, rather than being motivated to cooperate by their own payoff (Ariely, Bracha, & Meier, 2009). It appears that the presence of these incentives can 'crowd-out' cooperation in public settings (Kraft-Todd et al., 2015).

Reputation information must be communicable to be effective

One key assumption of the above research, suggesting that the opportunity to gain a good reputation can inspire cooperation, is that this information must be communicable to potential interaction partners in some form (Rand & Nowak, 2013). In the social world, our interactions can take many forms, including interacting with strangers in online environments, where information about others is often conveyed across formalised reputation systems. In online market places, like eBay, these reputation systems have economic consequences, as vendors with good reputations are trusted more by buyers and can thus stand to earn more than those where no reputation is available (Resnick, Zeckhauser, Swanson, & Lockwood, 2006).

Similarly, information about others can be transmitted verbally across actors in social networks. For example, gossip about others' behaviour affects how people interact with that person and cooperation is higher when people encounter positive gossip about an agent (Sommerfeld, Krambeck, Semmann, & Milinski, 2007). However, human memory is imperfect, so gossip may sometimes convey inaccurate information about the reputation of potential social partners. Ultimately, direct 'behavioural experience' with different social

partners is important and affects the social judgements that we make about interaction partners (Bayliss & Tipper, 2005; Heerey & Velani, 2010; Shore & Heerey, 2011). These judgements may also affect the decision to cooperate with that partner (Hoffman, Yoeli, & Nowak, 2015). Part of this behavioural experience with social partners therefore comes from observations of their behaviour and the ability to receive social cues from these partners.

Social cues can signal intentions for cooperation

Indeed, observable behaviour can signal the intent to be cooperative (Jordan, Hoffman, Nowak, & Rand, 2016). Also, individual social cues can also be influential on decisions to cooperate. For example, research has suggested that positive social signals increase cooperation (Scharlemann, Eckel, Kacelnik, & Wilson, 2001) and that emotional expressivity in general is predictive of cooperative intent (Schug, Matsumoto, Horita, Yamagishi, & Bonnet, 2010). Moreover, some social cues, i.e., genuine smiles, are sent and interpreted as honest signals for trustworthiness and yield higher payoffs for dyads (Centorrino, Djemai, Hopfensitz, Milinski, & Seabright, 2015).

Similarly, research has found that initiating laughter within a dyad is correlated with increased levels of cooperation, suggesting that this cue may also be an honest signal for cooperative intent (Burton-Chellew & West, 2012). This finding comes from studying a high-stakes situation in which realistic social interaction was possible – on a televised British gameshow "Golden Balls". Here, contestants must decide whether to 'split' or 'steal' an accumulated pot of money, which in some cases can be substantial, in a variant of the Prisoners Dilemma game (van den Assem, van Dolder, & Thaler, 2012)[7]. Interestingly,

[7] See van den Assem et al., (2012) study details for a particularly thorough explanation of the gameshow procedure. Although these studies bring with them their own unique caveats (such as the self- and commercial-selection of participants, for example), they also provide a novel opportunity to study decision-making under conditions that would be difficult to replicate in the laboratory. For example, van den Assem et al., (2012) estimated that it would cost £2.8m to replicate the Golden Balls setup under experimental conditions. As one of the oft-cited limitations to experimental games are that they are low-stakes (see List & Levitt (2005) for a full

research has also found that the way in which contestants spoke to each other and made promises was also indicative of their final decision. That being, players who used explicit and unconditional statements such as "I will split" were more cooperative than players who used more malleable, conditional statements such as "I will split if you split" (Turmunkh, van den Assem, & van Dolder, forthcoming 2019). This would suggest that it is possible to infer how trustworthy a player is by the way they speak.

Back in the experimental world, research has found that judgements of a partner's trustworthiness can also lead to increased cooperation with that partner, even if this signal is misleading (van 't Wout & Sanfey, 2008). Thus it seems that positive social signals/social judgements of interaction partners, can cause people to overestimate the value of these signals (Averbeck & Duchaine, 2009; Gaertig, Moser, Alguacil, & Ruz, 2012; Shore & Heerey, 2011), leading to increased trust/cooperation with that partner, even if their behaviour is actually no different to that of more 'negative' social partners (Ruz, Moser, Webster, McCandliss, & Quartz, 2011).

The role of face-to-face communication with social partners

An early review paper on the role of communication in economic games suggests that the opportunity to communicate with a game partner helps to increase cooperation compared to when this opportunity is absent (Sally, 1995). Interestingly, the opportunity to communicate with others may present the opportunity to negotiate with partners to elicit cooperation, even if this occurs via written communication rather than face-to-face interactions (Bochet, Page, & Putterman, 2006). However, it appears that in general, cooperation is better sustained via face-to-face interactions than through written

---

discussion), these studies can give us a glimpse of the social processes that may underpin decision-making under tangibly high-stakes situations, helping to address such caveat.

communications (Balliet, 2009) and these opportunities also mean that people are more honest (Van Zant & Kray, 2014). It also seems that whilst face-to-face interaction makes people more cooperative, this effect is mediated by the level of rapport between the dyad (Drolet & Morris, 2000). Thus, it seems that it is not merely the ability to interact with others that fosters cooperation but the social processes occurring between partners that also play a role in decision–making (Bicchieri & Lev-On, 2007).

It is also important to note that studies with communication opportunities often involve discussion of *intentions,* examining the role of this type of communication plays in fostering cooperation specifically (e.g., Arechar, Dreber, Fudenberg, & Rand, 2017; Isaac & Walker, 1988). However, these interactions are qualitatively different to interactions that do not involve explicit negotiation (Putnam & Jones, 1982), and may actually engage fairness norms in explicit ways that are atypical of natural social processes (Welsh, 2004).

As such, there are very few studies that actually allow for social interaction during the decision-making process; much less are those allowing for *naturalistic* interactions similar to those in everyday social exchanges. An early study that did allow for some element of free-form discussion between participants hypothesised that this would have no bearing on participants' decisions, as the economic model prediction for games with communication were the same as those without (Ostrom, Walker, & Gardner, 1992). However, in this study, earnings from games allowing for communication between players (that did not include discussion of game strategies) were in fact substantially larger than when no communication was allowed. This was an early indication that players may actually take into account factors from the social environment when making economic decisions.

Gächter & Fehr (1999) further contributed to this idea by comparing cooperation in the PGG under different 'social exchange' conditions. Players could either interact face-to-face before and/or after the game, or participate in an anonymous control condition. 'Social

exchanges' taking place before the game were designed to establish minimal familiarity between participants. Exchanges after the game involved the group analysing each player's contribution decisions throughout the game. Findings indicated that contributions from the condition where players could interact before *and* after the game were larger than all other conditions. Contributions in all conditions decayed over time, except for those on the final round of the game (of which players knew about in advance) where all contributions increased, except for those in the anonymous condition. These results suggested that establishing some familiarity between participants pre-game and the possibility of interacting with them again at the end of the game was enough of an incentive to boost contributions.

One observation of the above study is that the pre-game social interactions appear stilted. Players were given topics to discuss (study topic, hobbies) and then were required to play a 'guessing game' with each other. The latter seems particularly unrepresentative of real-world interactions upon meeting someone new. Additionally, although players could have post-game discussions about contribution decisions, they were not able to communicate throughout the game. However, in the real world, we are often engaged in long term relationships and alter our decisions based on the most up to date information we have on interaction partners, gleaned from social interactions.

This issue is also relevant to a more contemporary study that either allowed group members to either interact face-to-face, via a chatroom, or to exchange 'cheap talk' (promises about their future contributions) before engaging in a PGG (Bochet et al., 2006). While this study found that pre-game face-to-face discussion was effective in establishing larger contributions (96% of the maximum possible contribution) compared to the chatroom condition (81% of maximum), it did not allow for continuous interactions across the game period, nor did it consider or record the quality of these social interactions. This is noteworthy because much of the theory about factors that influence social behaviour come

from ‘pseudo-social’ studies that can only approximate real world social stimuli; meaning that evidence of the characteristics of effective *real world* social interactions largely remain unknown and understudied (Heerey, 2015).

As a result, in this thesis, we attempt to capture and decode elements of naturalistic social interactions occurring throughout the process of cooperative decision-making, contrasted against typical anonymous laboratory settings. This therefore provides a first look at the influential characteristics of these group level social interactions, discussed Chapters 2 & 3.

Social versus monetary incentives for cooperation

In the first two chapters, we were also interested in examining the effect of different types of punishments and rewards across the aforementioned social interaction contexts (i.e. face-to-face versus typical anonymous laboratory settings). We were interested in exploring how ‘social’ incentives may fare compared to monetary. We designed our punishments (and rewards) based on methodological observations from two key papers – Masclet, Noussair, Tucker, & Villeval (2003); Noussair & Tucker (2005). These papers were amongst the first to theorise about the role of ‘informal sanctions’ in cooperative decision-making, i.e., social disapproval, versus the typical ‘formal sanctions’ (i.e., monetary punishment) often studied in laboratory economics games (Masclet et al., 2003).

In these papers, groups of four players engaged in a multi-round PGG in an anonymous laboratory environment – that being one in which players could not see or interact with group members face-to-face, but instead played via computer terminals. After each round, contribution amounts were displayed on a player's screen (not explicitly tied to the contributor's identity) and participants could then decide whether they wanted to punish someone. In the monetary punishment condition, players could anonymously assign 0-10 punishment points that would lower the recipient's return by an increasing amount as defined by a punishment schedule. In the non-monetary punishment condition, informal sanctions were operationalised as 'disapproval points', whereby a player could assign 0-10 disapproval points to indicate how much they disapproved of another player's contribution behaviour on that round. This non-monetary version of the punishment was designed to be directly comparable to the punishment points in the monetary condition. In both studies, results indicated that the availability of monetary sanctions lead to higher average group contributions than non-monetary sanctions. Results from Noussair & Tucker (2005) highlight this effect especially well in Figure 1.1.

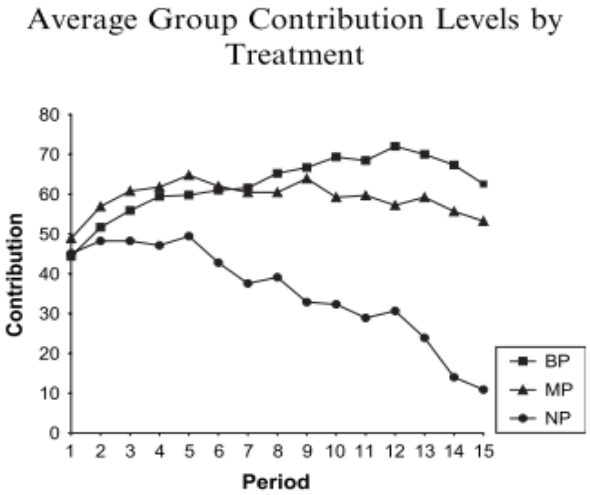


Figure 1.1. Results figure taken from Noussair & Tucker (2005), displaying the decay of average group contribution over experimental period in the non-monetary (NP) condition, i.e., the use of disapproval points as an informal sanction. Monetary punishments (MP) appeared to sustain average contributions over time, as well as when both punishments were available (BP).

Although these papers were the first attempt to incorporate the idea that social factors may play a role in cooperative decision making, we observe several methodological issues here. Firstly, it is unusual for social partners in the real world to express disapproval by assigning a number of 'points' to someone. This appears to be very abstract version of social punishment to that which the authors intended to study, and so it is unclear what these results may be telling us about how this punishment type may shape cooperative behaviour. Secondly, the context in which players interacted in these studies is also an atypical environment compared to that which participants would normally experience (as discussed at length in: Levitt & List, 2007) in that real world interactions are rarely completely anonymous. And so, it is unclear how these results would differ if social punishments and interaction contexts were manipulated to be more representative of real life; that is if participants were allowed to directly communicate their disapproval (or any other emotional expression) towards those receiving a punishment.

Accordingly, we explore whether cooperation in the Public Goods Game is affected by punishment type and social interaction context in Chapter 2. Additionally, research has found that although the availability of punishment results in larger contributions to the public good compared to when it is not, individual payoff is actually higher when rewards are available compared to punishment (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). Therefore, in Chapter 3, we also explore the effects of social versus monetary rewards on cooperation over different social interaction context in Chapter 3.

The importance of social feedback

We expect that the transmission of social information, in face-to-face interaction settings, in response to punishments and rewards, may affect future economic decisions. Thus, we expect that social feedback from interaction partners to be important in shaping other forms of decision-making. Accordingly, we also expect social feedback to affect the

process of learning on an intrapersonal level. Indeed, social cues are abundant in the environment and so, the ability to learn from other agents in our environment is important, especially as interactions change dynamically on a moment-to-moment basis (Kringelbach & Rolls, 2003).

Being attuned to social cues from interaction partners is an important evolutionary function as these cues convey important information about the environment, as well as the sender's emotional and mental state (Emery, 2000). Individuals can then use these cues from others to predict what they will do in the future (Heerey & Velani, 2010). Thus, it also follows that information from others may also direct an individual's own behaviour; to inform one of what they have done well, or what behaviours in which one should cease to engage. Indeed, subtle changes in social cues from others are used to dynamically adjust an individual's behaviour (Kringelbach & Rolls, 2003); to maximise the chance of reward and minimise that of punishment (Heerey, 2014). Furthermore, some individual social cues, such as smiles, are intrinsically rewarding (Averbeck & Duchaine, 2009; Shore & Heerey, 2011), and these social rewards are more valuable compared to non-social alternatives, even when these stimuli convey the same objective value (Heerey, 2014; Shore & Heerey, 2011). Thus, the use of social stimuli as feedback is more effective compared to non-social, symbolic feedback during an associative learning task (Hurlemann et al., 2010).

However, an avenue in which the effect of social feedback has not been explored, is during implicit learning tasks. The process of implicit learning is interesting as we often learn abstract rules about the world without explicit instruction (Janacsek & Nemeth, 2012), thus, implicit learning can be defined as the process of learning a contingency without explicit awareness (Shanks, 2005). An effective method for measuring the process of implicit learning, beyond surface level indices, is the Event Related Potential technique (Baldwin & Kutas, 1997; Eimer, Goschke, Schlaghecken, & Stürmer, 1996). Currently, we do not know

whether the type of feedback (i.e., social versus non-social) affects the process, or neural signature, of implicit learning. Thus, in the final empirical chapter of this thesis (Chapter 4), we are interested to explore whether exposure to social feedback, especially social rewards (i.e., smiles; Heerey & Crossley, 2013; Shore & Heerey, 2011), compared to non-social 'symbolic' feedback, differentially affects the process of implicit learning during a novel card game.

Preface

In the following empirical chapters, I thus explore the effect of social information, including rewards and punishments, during cooperative decision-making and implicit learning. It is important to note that throughout this thesis, I will be studying proximate explanations of social behaviour and cooperation. These explanations are those that are concerned what a behaviour looks like rather than why it exists (i.e., an ultimate explanation). For example, cooperation may be explained by individual predispositions to reward/punish others (Fehr & Fischbacher, 2003), or by experiencing 'moral emotions' (e.g. guilt), that regulates altruism to allow cooperation between social partners (Trivers, 1971). As opposed to providing answers to the question of ultimate cause – i.e., *why* cooperation between individuals might exist, this thesis instead contributes to the literature that provides proximate explanations of behaviour that underpins cooperation.

In Chapter 2, I address the question of how naturalistic social interactions may affect cooperative behaviour during the Public Goods Game, an aspect of decision-making that has so far been overlooked in this field. Furthermore, in this chapter, I compare different types of punishments (monetary, social or both) to further examine whether the efficacy of punishment type depends on the social setting in which it is administered. In Chapter 3, I extend this novel variant of the Public Goods Game to examine the effect of different types of rewards on cooperative decision-making. Finally, in Chapter 4, to explore the potential

facilitatory effect of social rewards on learning, I compare social and non-social stimuli used as feedback during a novel implicit learning paradigm.

Chapter 2

The social and economic costs of punishment: Evidence from a Public Goods Game

Philippa J Beston

Erin A Heerey

Author contributions: P.J.B and E.A.H designed experiments and collected data. P.J.B & E.A.H co-designed and implemented the 'investment behaviour', 'cost of punishment' and 'sensitivity/positivity' analyses. P.J.B wrote the paper. E.A.H proofread and suggested edits.

Abstract

Cooperation in economic games is a puzzle. Under certain laboratory conditions, this strategy has the potential to reduce individual payoffs, thus is not economically rational. Yet, cooperation is widely observed in such games, particularly when the option to punish free-riders is present. Research in laboratory settings suggests monetary punishments are superior in maintaining cooperation to non-monetary "social" alternatives (Masclet et al., 2003; Noussair & Tucker, 2005). In contrast, there is also a large body of literature suggesting that reputation incentives maintain cooperation. Here, we investigated these ideas by comparing a standard, anonymised Public Goods game with a socially enhanced version of the game in which participants could freely interact with each other. We also compared monetary (return reduced by 50%), social (lowest contributor publicly named) and combined punishment versions. In the anonymised game context, we replicated common findings: monetary punishments improve cooperation compared to social punishments. However, this pattern reversed in the face-to-face setting. Although in our game punishments were not costly to distribute, we did find that larger costs were associated with social punishments in this context, but punishments in the anonymised context did not appear to be as unpleasant to give/receive. Furthermore, In the face-to-face condition, we found that group interaction positivity predicted players' contributions on the next round in conditions involving social punishments. Together, these results suggest that access to social information and the immediacy of reputation incentives drove cooperative decision-making in the face-to-face game in a way that was not present in the anonymised game.

Introduction

Every day, people contribute time, effort and resources to support public goods - individually subscribed provisions designed for public benefit. These provisions add value to people's lives by ensuring access to goods and services and evening resource distribution across individuals (Gravelle & Rees, 2004). Public goods operate at both global and local levels, including international funds for disaster or humanitarian relief, contributions to nationalized health care programs and charitable donations. Because of their reach and consequence, understanding how and why people contribute is important for both policy makers and social partnerships that rely on personal public contributions (Parks, Joireman, & Van Lange, 2013; Rand, Yoeli, et al., 2014).

Unfortunately, public goods provisions are often costly to run and contributors bear financial or effort-related burdens to maintain these resources. They are also 'non-excludable', meaning that anyone can access them, including those who have not contributed their fair share. This presents the opportunity to "free ride" on the cooperation of others (Isaac, McCue, & Plott, 1985; Kim & Walker, 1984), thereby creating a social dilemma in which those who invest highly subsidise other players' benefits (Isaac, Walker, & Williams, 1994; Marwell & Ames, 1980).

To explore the landscape of public goods provision, experimental economists have developed the public goods game. In basic games, participants receive an endowment (usually in money or tokens) and independently choose how much they wish to contribute to a public fund. The fund is then tallied, multiplied by some factor (typically greater than one and smaller than the number of players) and the result split evenly amongst the players, regardless of initial contributions (Andreoni, 1995).

Although a contribution of zero is the dominant strategy in the public goods game (Olson, 1965, as cited in Ostrom, 2000), evidence shows that players are often more cooperative than "rational" economic models would predict (Camerer & Fehr, 2006; Kahneman

& Tversky, 1979; Roth et al., 1991), showing a willingness to sacrifice their own personal payoffs in favour of the group's return (Isaac & Walker, 1988; Isaac et al., 1994; Rand, Dreber, Ellingsen, Fundenberg, & Nowak, 2009). One explanation for this behaviour is that players are continuing to employ the cooperative intuitions that are fruitful in everyday social interactions (Rand et al., 2014; Rand, Greene, & Nowak, 2012). Indeed, cooperation may be the "default" state due to repeated interactions with social group members, (Andreoni, 1988). In the laboratory, however, players tend not to maintain initial cooperation levels, as they learn that unreciprocated generosity does not, on average, yield profitable results (Andreoni & Croson, 2008; Burton-Chellew & West, 2013; Burton-Chellew, Nax, & West, 2015; Croson, 1996; Fehr & Gächter, 2000). Moreover, in laboratory settings, there are often negligible real-world repercussions following defection (Guala, 2012).

Interestingly, the inclusion of social information may promote generosity. For example, people are more likely to be generous to those who have been generous with others (Fischbacher, Gächter, & Fehr, 2001; Wedekind & Milinski, 2000; Seinen & Schram, 2006). Moreover, when people's actions are directly observable, behaviour changes for the better (Andreoni & Petrie, 2004; Kraft-Todd et al., 2015). This suggests that people have additional motivation to cooperate when they can gain a prosocial reputation (Ariely et al., 2009; Engelmann & Fischbacher, 2009), potentially to signal their trustworthiness to future interaction partners (Jordan, Hoffman, Nowak, & Rand, 2016). However, when behaviour is not directly observable, defection becomes prevalent (Hoffman et al., 1994).

Another effective method of maintaining public goods contributions is the introduction of monetary sanctions (Balliet et al., 2011; Fehr & Fischbacher, 2004a; Henrich et al., 2006). Indeed, evidence shows that when contributions deviate from expectations, people are prepared to punish the perpetrators, even when this is a costly option (Fehr & Fischbacher, 2004b; Fehr & Gächter, 2000a; Herrmann et al., 2008). To implement targeted sanctions, participants

typically receive information about contribution amounts, even though this information is not explicitly linked to contributor identity (Fehr & Gächter, 2000, 2002; Masclet, Noussair, Tucker, & Villeval, 2003; Noussair & Tucker, 2005; Rand et al., 2009). However, this practice presents a potential problem. Although it allows for targeted punishments, it also provides a 'descriptive norm' (Cialdini, 2003) about how others are playing, which may alter subsequent contributions. Thus, it is it difficult to disentangle the effects of costly punishment from the effects of normative information on participants' economic decisions.

Research has sought to compare the effects of monetary punishments with sanctions designed to emulate social disapproval, operationalised, for example, by assigning disapproval 'points' to defecting players (Masclet et al., 2003; Noussair & Tucker, 2005). Interestingly, results from these studies suggest that non-monetary sanctions seem to be less effective than monetary sanctions in promoting cooperation in the laboratory. In field studies, however, interventions manipulating an individual's social concerns appear to be consistently effective at maintaining cooperation, compared to cost/benefit manipulations (Kraft-Todd et al., 2015).

One reason for this discrepancy may be that in the real social world, people directly communicate disapprobation (Gächter & Fehr, 1999), thereby enhancing its effectiveness. In this study, for example, participants who had face-to-face interactions before and after the game contributed significantly more than those who did not have interaction opportunities (Gächter & Fehr, 1999). Other work also shows face-to-face interaction to be beneficial for cooperation (Balliet, 2009; Drolet & Morris, 2000; Ostrom, 2000) and honesty between social partners (Van Zant & Kray, 2014). Thus, in socially impoverished laboratory contexts, where interaction with social partners is not possible, the "sting" of social disapproval may be reduced, as there are no tangible social consequences for punished players (Masclet et al., 2003; Noussair & Tucker, 2005). However, it is so far unknown how face-to-face social

interaction and more naturalistic social punishment may interact to influence cooperative behaviour.

Here, we resolve these controversies by asking two questions. First, we use a multi-round public goods game to ask how much people change their behaviour when receiving punishments in the absence of normative information (i.e. the contribution amounts of other group members). Second, we ask how the game context (social 'face-to-face' or non-social 'anonymous') changes the relative effectiveness of monetary versus social sanctions. Assuming that normative information is not solely responsible for maintaining contribution levels, we predicted that in the typical anonymous version of the game, monetary, but not social punishment would serve to maintain contribution levels, thereby replicating previous findings.

Experiment 1

In this experiment, we implemented a standard version of the multi-round public goods game using an internet-based protocol. This protocol is similar to many previous experiments in that players experience non-manipulated games in a fully anonymous context (e.g., Andreoni, 1988) and it includes both reputational (labelled 'Social' punishments for simplicity) and Monetary punishments. Here however, we examine punishments in the absence of descriptive data about others' investments by making the punishments "free" to administer, with the caveat that the lowest contributor is always punished (including, potentially, the punisher). We can therefore examine how much extra money people are naturally willing to contribute in order to avoid punishment, how much punishment experience changes future contributions in the absence of normative data, and how social and monetary punishments compare on these factors.

Method

Participants

One hundred twenty undergraduate participants completed a public goods game in exchange for partial course credit and a small monetary bonus. The sample consisted of 94

females and 26 males (mean age=19.81; SD=2.60). Participants provided written informed consent before participating and the University's Ethics Committee approved all study procedures (likewise for Experiment 2 below). The sample size was determined in advance based on both budgetary limitations and the number of groups we estimated we would be able to recruit over the course of the year. Data analysis began only after data collection was complete (likewise for Experiment 2).

Procedure

Participants attended the experiment in groups of four for an iterated 15-round public goods game, played online via networked computer terminals. They arrived to a student lounge outside a busy campus computer lab. The experimenter greeted them individually and showed them to computers in different sections of the lab. Importantly, the lab was in full operation during the experiment, meaning that players were not aware of which other lab users were playing the game or which randomly assigned colour each player played.

A purpose-built website (controlled by a MySQL database) coordinated the game, firstly providing players with standardised instructions (Appendix A), then allowing players to make their contributions independently (Figure 2.1a), whilst viewing round feedback simultaneously. Players could not move to the next round before all four investments were recorded (Figure 2.1b)

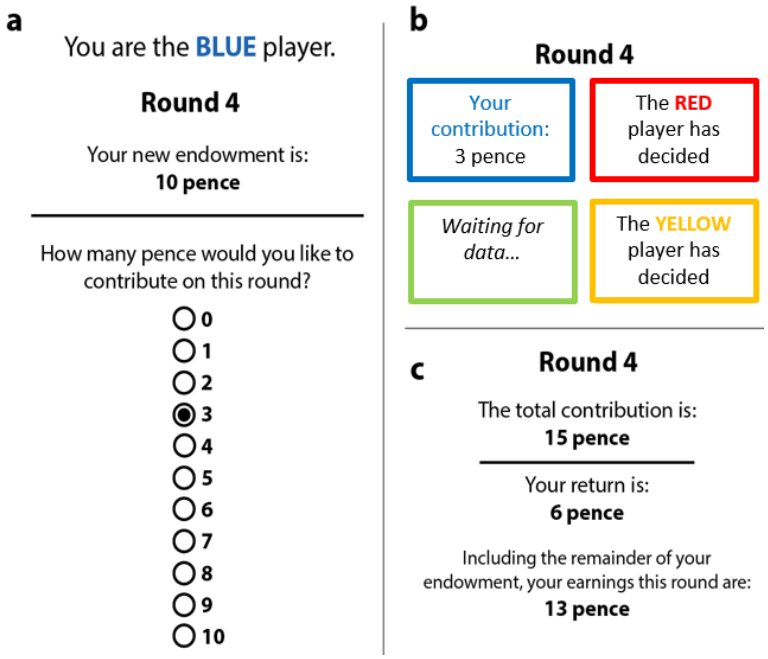


Figure 2.1. Web protocol. Example of a non-punishment round. a) Page on which participants indicated their contributions; b) contributions as received; c) round feedback.

and investment feedback displayed for a minimum of 2 seconds (Figure 2.1c). The contribution phase began with an endowment of 10 pence (0.10GBP). Participants then chose the amount they wished to contribute to a "group resource." They retained the rest in a private "bank." Participants made their contributions by clicking radio buttons on the webpage. After the contribution phase, the database tallied the total fund, multiplied it by 1.6 (Zelmer, 2003) and calculated participants' returns. The return was 25% of the total fund. If the total fund was not evenly divisible by 4, it was rounded up to nearest integer so that each player received the same return. This methodological feature was necessary because in Experiment 2, we used endowments of real pennies (GBP 0.01). Thus, we were not able to give returns from the group contribution in denominations smaller than 1 penny. We therefore used the equivalent experimental currency in this version of the game. Participants then received feedback about the group's total investment and their individual return (Figure 2.1c). Players were unaware of the number of rounds they would complete in the game. However, they were aware that they played the same group of participants throughout all game rounds.

Participants played five "practise" rounds (with no punishment options available). After Round 3, they completed a quiz assessing their understanding of the game (Appendix B(i). See also Appendix B(ii) for details on comprehension scoring and analysis). Participants always received a standard reminder about the "rational" contribution strategy after the quiz that encouraged them to consider their returns in the context of their contributions, and in the context of other players' possible strategies (Appendix C).

After five practise rounds, the website introduced the opportunity to punish free-riding players. This task phase began with instructions about the protocol for punishments. Beginning on the contribution page for Round 6, each player viewed a set of four coloured "punishment tokens" (small squares matching the player's colour). If a player decided to punish another, he/she ticked a button on the screen and one of the available tokens

disappeared. Players made this decision on the same screen as they made their contribution decision and were aware that this punishment would be applied based on their investment behaviour on that *current* round. The database applied the punishment to the player who had made the lowest contribution to the group's total on that round, randomly selecting amongst equally low contributors.

We opted to use punishment in this way, as opposed to allowing participants to target their punishments based on how other players had contributed (as in Rand, Dreber, Ellingsen, Fudenberg, et al., 2009, for example), so that we could determine the effect of punishment in the absence of specific normative information about others' contribution strategies. Evidence suggests that such information may influence contribution decisions in a large proportion of players (Fischbacher et al., 2001), and yet many studies display players' (anonymised) contribution information when allowing players to choose a punishment target (Masclet et al., 2003; Noussair & Tucker, 2005; Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). Therefore, by removing access to contribution information, we can determine how players alter their contributions when giving/receiving punishments when they do not have the ability to make direct conclusions based on their knowledge of how other players' contributions vary over rounds. Note that this feature also meant that punishments were democratically distributed to the lowest contributor, thereby limiting the ability for players to use punishments 'antisocially', in which players enforce a particular level of contribution by punishing those who are seen to be over-contributing (Herrmann et al., 2008).

To indicate that a player had been the lowest contributor, he/she received a black punishment token and text feedback with return information (e.g., "On this round, YOU have received a punishment!"). This meant that the identity of the *punisher* remained anonymous. Players received only four punishment tokens for the 10 game rounds in which punishment was possible. These were not replenished if a player chose to distribute them all. If more than one

player chose to punish on any round, the extra punishment was held in reserve and applied on the next round in which no player had chosen to punish. Participants were made aware of this contingency when they received instructions for the punishment rounds. This feature was to allow there to be a peer-implemented punishment on most of the rounds of the game, as peer-implemented (or 'decentralised') punishment appears to be a stronger moderator of cooperation than 'centralised' punishments (i.e., those that are determined by the experimenter or external authority; Balliet et al., 2011). Furthermore, this feature also meant that the threat of punishment would be omnipresent, much like it is in many real world social interactions. As a small scale example, those who neglect their duties in a shared household run the risk of receiving hostility from a disgruntled housemate or landlord at some future point, even if they avoid 'punishment' immediately. Thus, by allowing unused punishments to carry over to future rounds, we are able to imitate this looming threat of punishment for antisocial behaviour, adding an element of realism to our game.

There were three punishment conditions: monetary only; social only; and monetary and social punishments combined. Punishment type was a randomly assigned, between-groups manipulation (10 groups per condition), so all members of a group experienced only one type of punishment. Importantly, because the database randomly assigned punishment conditions to groups, the experimenter was blind to which punishment condition was active within a group.

In the monetary-only punishment condition, punished players received a reduced return (e.g., "The GROUP return is [12] pence. YOUR return is [6] pence."). The reduced return was 50% of the group return. If this number was not evenly divisible, it was rounded down to the nearest whole number. Importantly, only the punished player knew about the punishment. Therefore, the identity of the punished player remained anonymous.

In the social punishment condition, punished players received the same return as the other group members; however, the computer revealed the colour identity of the punished

player to all the players (e.g., "On this round, the RED player has received a punishment!"). Thus, although there was no financial penalty in this condition, the colour identity of the lowest contributor was known to the group. This allowed us to examine the effect of reputation building and maintenance on contribution levels. The third punishment condition included both social and monetary punishment, meaning that both punishments (publication of the punished player's colour along with a reduced return) were applied.

Finally, in contrast to many public goods games in which punishment is possible, our players did not "pay" in order to apply a punishment (i.e., these were not "costly" punishments per se). Instead, the lowest contributor received the punishment. This meant that if the player who chose to punish was *also* the lowest contributor, he/she was the person who received the punishment. Because players knew about this contingency in advance, we were able to examine the extent to which players *naturally* increased their contributions on rounds in which they included a punishment token, rather than enforcing a fixed cost as defined by the experimenter (as in, for example: Fehr & Gächter, 2000; Masclet et al., 2003; Noussair & Tucker, 2005). This therefore allows us to determine whether different punishment types (Monetary, Social, Combined) are particularly costly, and whether this cost differs across different social contexts (anonymous, face-to-face). Participants played 10 rounds of the game with punishment options available. After this, the game ended and participants were debriefed, paid their game earnings (approximately £2.25), and dismissed.

Data Analysis

Investment behaviour. We analysed average participant investment on punishment trials 6-15 using a one-way ANOVA model, implemented in SPSS.

Cost of Punishment. Usually, the cost for a participant to punish another player is defined by the experimenter (Masclet et al., 2003; Noussair & Tucker, 2005). Here, we were

interested in determining whether the 'natural' cost that players were willing to pay differed depending on punishment type.

To determine the cost that participants were willing to bear to administer punishment (here named the "cost to give" a punishment), we calculated the average contribution for each trial in which a player administered a punishment during trials 6-15 (i.e., the punishment phase). We then compared this to that player's average contribution, also on trials 6-15, in which the option to punish was available, but where that player had neither included a punishment, nor received one on the previous round – subsequently refered to as a player's **"standard rate"** contribution. Thus, the "cost to give" punishment variable was calculated as the difference between an individual's average standard rate contribution and their average contribution on trials where they chose to give a punishment, excluding trials immediately after punishment receipt (as contributions may vary because of punishment receipt). We also used this same calculation for the trial *after* a player received a punishment to give a difference score from a player's standard rate contribution, called the "cost to receive" a punishment. This gives an indication of the average change in contribution, compared to the average standard rate, on the trial after a player received a punishment. This analysis therefore provides a metric of the degree to which a punishment is unpleasant. If punishment is effective at enhancing contributions, a punished player's next contribution after punishment receipt should be significantly higher than his/her typical, pre-punishment contribution.

Trials in which a player received a punishment in the previous round and chose to punish on the next, (i.e., punishments that could have been motivated by the desire to retaliate) were removed from this analysis. On the trials after receiving a punishment, players chose to retaliate around 42% of the time in Experiment 1 and 38% of the time in Experiment 2. Retaliatory punishments also constituted about 30% of the total number of punishments

distributed in Experiment 1 (Monetary: 32 retaliations of 113 punishments given = 28%; Social: 28/103 = 27%; Combined: 33/112 = 29%) and around 27% of trials in Experiment 2 (Monetary: 32 retaliations over 123 punishments = 26%; Social: 23/81 = 28%; Combined: 35/120 = 29%). Removing retaliatory trials meant that this analysis reflects only how much players changed their contributions to avoid their own punishments, eliminating the possibility that they may have also changed contributions when retaliating.

Additionally, it is important to note that trials in which a player gave a punishment and received his/her own punishment back (i.e., self-punishments) were included in this analysis. Specifically, self-punishments represent equivalent 'learning opportunities' with respect to the current cooperation level in the game as do other-punishment trials. Such outcomes are therefore equally valuable in terms of allowing players to correct their contribution levels.

We used a customised mixed-model, nested design ANOVA (in which individual players were nested within groups, adapted from the Social Relations Model; Back & Kenny, 2010; Kenny & La Voie, 1984) implemented in SPSS, to examine these difference scores (likewise for Experiment 2 and the analogous analysis in Chapter 3). Cost type (for Giving and Receiving punishment) served as the within-participants variable. Punishment Condition (Monetary, Social, Combined) was the between-participants factor. The dependent variable was the change in investment (pence) when giving/receiving punishment versus a participant's "**standard rate**" contribution – that being average investment on trials where a player has not given a punishment, nor did they received one on the previous trial. Post-hoc comparisons are Bonferroni corrected throughout the report, and corrected *p*-values are reported.

One other important note is that in no analysis have we made direct comparisons between the pre-punishment (trials 1-5) and the punishment (trials 6-15) phases. We have only analysed the pre-punishment phase to compare average investment across punishment groups. This was to ensure that groups were behaving similarly prior to the introduction of punishment

– i.e., that there had been no administration error or experimenter bias that systematically affected the groups before the punishment phase began.

Results and Discussion

Investment behaviour

Participants' investments across trials appear in Figure 2.2a. Broadly, these data show that introducing the option to punish increases contributions, as seen in previous research (Fehr & Gächter, 2002; Masclet et al., 2003; Noussair & Tucker, 2005). These data also show that the type of punishment is important, $F(2,116)=6.27$, $p=.003$, $\eta_p^2=.10$. Unsurprisingly, we replicate previous results showing that "social"/reputational punishments are less effective at maintaining high contribution levels than punishments with a monetary component in anonymised, laboratory environments (corrected p-values<.048). There were no differences between monetary punishment only and combined punishment conditions ($p=.937$). The punishment conditions did not differ in contribution levels prior to the introduction of punishment, $F(2,116)=0.30$, $p=.743$, $\eta_p^2=.01$.

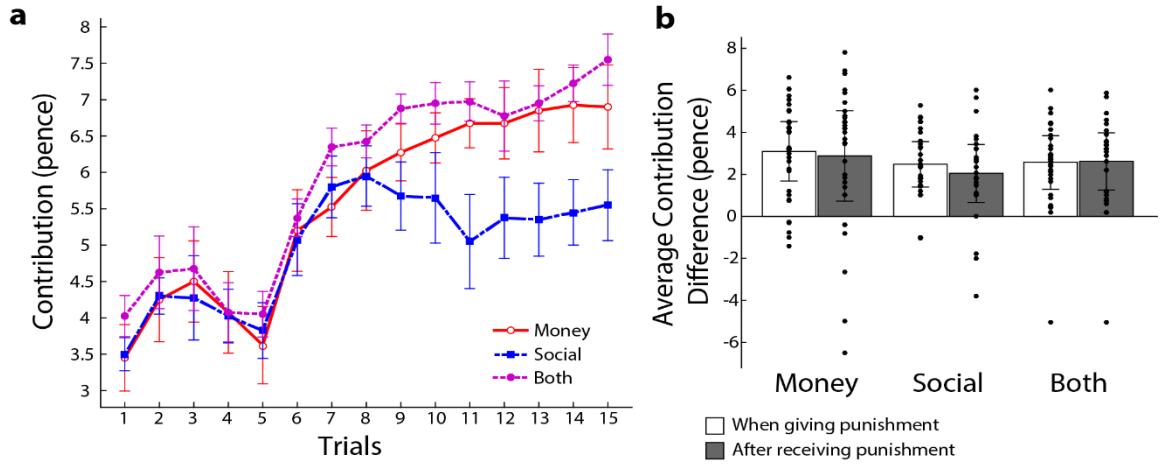


Figure 2.2. Experiment 1 results. a) Contributions by punishment condition across trials. *NB*. This figure gives a descriptive representation of average individual contributions by round number, per condition, only. For statistical analysis, we analysed average contribution by punishment condition, collapsed across trials 6-15. b) Average increase in contributions when giving (white bars) and after receiving punishments (grey bars) by punishment condition. Individual data points are superimposed on the group averages. Error bars show the 95%CIs.

Punishment descriptive statistics

When available, participants typically opted to use most of their punishment tokens, with 18.3% of players using 3 (of 4) tokens, and 46.7% using all 4 tokens. This therefore meant that most trials included the punishment of another player – 85% of Monetary trials, 76% of Social, and 89% of Combined trials. However, statistically, there were no differences in the number of punishments administered across the punishment conditions, $F(2,117)=0.35$, $p=.708$, $\eta_p^2=.01$ (average number of punishments dispensed, per individual player: Monetary=2.83 (SD=1.52); Social=2.57 (SD=1.48); Combined=2.80 (SD=1.49)). Of all the trials that contained a punishment, punishments rebounded onto the distributor (i.e., self-punishment) on 23.53% of monetary trials, 18.42% of social trials and 32.58% of combined trials.

Cost of punishment

Although both giving and receiving punishment led to increased contributions relative to pre-punishment trials (confirmed by one-sample *t*-tests: *t*-values>4.85, *p*-values<.001; Figure 2.2b), results showed that there were no statistically significant differences between giving versus receiving punishment, $F(1,78)=0.03$, $p=.859$, $\eta_p^2<.001$, or across the punishment conditions generally, $F(2,78)=0.49$, $p=.612$, $\eta_p^2=.01$.

We had predicted that for the monetary punishment conditions, the costs of receiving a punishment would be greater than for social punishment only, as previous research suggests that monetary punishments are more effective than social punishments (Masclet et al., 2003; Noussair & Tucker, 2005). However, we did not find this to be the case, as the Punishment Condition (monetary, social, combined) by Punishment Cost (giving, receiving) interaction was not statistically significant, $F(2,78)=2.37$, $p=.100$, $\eta_p^2=.06$. Thus, all participants increased their voluntary contributions when they chose to punish and after receiving punishment, although there were no differences across conditions.

Sensitivity to inequality model

We additionally used our data to examine the degree to which participants appeared to be sensitive to other group members' contributions. That is, we asked whether participants become aware of inequality between group member contributions in the absence of descriptive information (as evidence shows that they are aware of such discrepancies when normative information is present (Chaudhuri, 2010; Fehr & Gächter, 2000a)). Specifically, we examined how the discrepancy between one's own contribution and the average of group member contributions on trial t predicted contributions on trial $t + 1$. We used the ordinary least squares regression method (Hayashi, 2000), to estimate the following first order autoregressive model:

$$Y_{it+1} = C + \alpha(Y_{it} - \overline{Y}_i) + \beta\left(X_{jt} - Y_{it}\right) + \varepsilon_{it}$$

In this model, C is the constant and models the intercept (average investment). α is the autoregressive coefficient, which models linear dependency in a time series, (i.e., the extent to which a participant's contribution on each trial depends on his/her typical contribution strategy; all punishment conditions modelled together). Y_{it} is participant i's contribution on trial t. $\overline{Y}_i$ is the average of all participant i's other contributions. X_{jt} refers to the average contribution of the other players in group j on trial t, so the difference between X_{jt} and Y_{it} is the discrepancy between a player's contribution and the average contribution of other group members on that trial. β is the estimated regression coefficient for this term. This was modelled separately for each punishment condition. In order to compare the estimated terms in the model, we mean-centred the raw data, meaning that the estimated terms code deviations from the mean.

We now examine whether participants' sensitivity to contribution inequity shapes future investments. One way to examine how sensitive people are to other players is to examine the extent to which their next contributions depend on the current average contributions of their group members. For example, if a player contributes three pence of a 12-penny fund, then the

average of the other players' contributions is also three pence, meaning that the contributions are equitable (a discrepancy of zero). However, if a player contributes six pence of a 12-penny fund, then the other players have contributed an average of two pence, resulting in a significant discrepancy (four pence) between one's own and others' contributions.

To examine the degree to which players were sensitive to group members' contributions, we examined the strength of the estimated terms in our regression model. Table 2.1 shows these results. Results showed that after accounting for participants' own contribution strategies (e.g., a participant's general tendency to make high contributions), participants used the discrepancy between their own contributions and the average contribution of the other group members to determine their next contributions (regression coefficients greater than zero). This occurred even in the absence of descriptive information about others' investments. We did not find significant differences across punishment conditions in this effect.

Together, results of Experiment 1 broadly replicate previous research findings showing that monetary punishments are effective at enhancing contributions over time. However, whereas social punishments were immediately effective (contributions increased on the trial following punishment), they did not maintain contributions to the same degree as did the threat of monetary punishment. Interestingly, in economic terms, the unpleasantness of the three punishment types was similar in terms of how much participants enhanced their next

		Groups modelled together		
α (Autoregressive Component)	Regression Coefficient	0.71		
	95%CI: Lower Bound	0.64		
	95%CI: Upper Bound	0.79		
		Monetary Punishment	Social Punishment	Combined Punishment
β (Sensitivity to Inequality)	Regression Coefficient	0.22	0.15	0. 13
	95%CI: Lower Bound	0.12	0.06	0.02
	95%CI: Upper Bound	0.32	0.25	0.23

Table 2.1. Experiment 1 Regression Results (anonymous setting). Estimated unstandardised model coefficients, 95%CIs.

contributions after receiving a punishment and in terms of how much extra they were willing to pay to avoid a punishment that they themselves had administered.

Experiment 2

As previous research shows, monetary punishments seem to be critical in keeping public contributions high. However, in the real world, public goods are often regulated within highly social settings. That is, one sees social partners on a daily basis and may therefore have a strong interest in maintaining interaction and relationship quality. In this type of setting, reputational information may serve an extremely important regulatory function in social decision-making. Insofar as people value reputation amongst face-to-face peers, they should be willing to pay to maintain it. We therefore ask whether social punishments enhance contribution strategies in an enriched social setting and whether the natural costs of giving and receiving punishment differ across punishment types.

Method

Participants

We recruited 124 participants from an undergraduate student research pool. Participants received partial course credit and a small monetary bonus based on their earnings in the game. The sample consisted of 83 females and 41 males (mean age=19.82; SD=2.69).

Procedure

This version of the game used a similar procedure to that in Experiment 1, except that here, players were not anonymous. Rather, they played a 15-round public goods game face-to-face, using real pennies as monetary tokens. To capture social influences on investment behaviour, participants were seated around a table and clearly visible to their fellow group members. We recorded their interactions via two wall-mounted video cameras located opposite to the long edges of the table at which participants sat (Figure 2.3). Although the cameras captured participants' social interactions, they were positioned so that investment behaviour remained private.

In order to track each participant's behaviour during the game, players received a colour-identity that they used throughout all 15 rounds. They sat behind color-coded screens to conceal their contribution behaviour from the rest of the group. Pilot research showed that the screens did not interrupt social interactions. Participants received standardised instructions throughout the game. At the start of each round, the experimenter gave each player an endowment of 10 pennies. Participants then chose the amount they wished to contribute to the public fund. They made their contributions by placing pennies into opaque, color-coded boxes and passing these to the experimenter. After receiving each player's contribution box, the experimenter passed these to a "banker" (a second experimenter) located in an adjacent room.

The banker, who had no contact with the group, entered participants' contributions into a computer program that recorded the data, calculated the total fund and computed the group's return as in Experiment 1. The banker then placed each player's return (rounding up to the nearest whole penny if necessary), in 1-penny coins, into his or her contribution box and returned these to the experimenter, who informed participants of the total contribution and the return on the round (e.g., "The total was 15 pence and your return is 6 pence."). The experimenter then returned contribution boxes and asked players to place all monies from that round (return plus endowment remainder) into their individual banks. The experimenter then started the next round by passing a new endowment to each player. Players were unaware of the number of rounds they would complete.

While participants waited for the banker to tally their contributions and dispense their returns, they were allowed to converse freely. However, the experimenter informed them at the start of the game that explicit negotiation about game strategy/contributions (both verbal and nonverbal) was forbidden. The interactions ranged from 46 seconds to 189 seconds (Mean length=118.11, SD=22.50). There were no differences in average interaction length between the punishment conditions, $F(2,27)=1.92$; $p=.166$. Measures of interaction mood and quality were rated from videos of these interactions (see 'Video ratings' section below). Due to a technical error, no interactions were recorded for one of the groups (social punishment) so we recruited an extra group of participants in that condition to replace the missing video data. Analyses of video data only necessarily exclude the group with missing data.

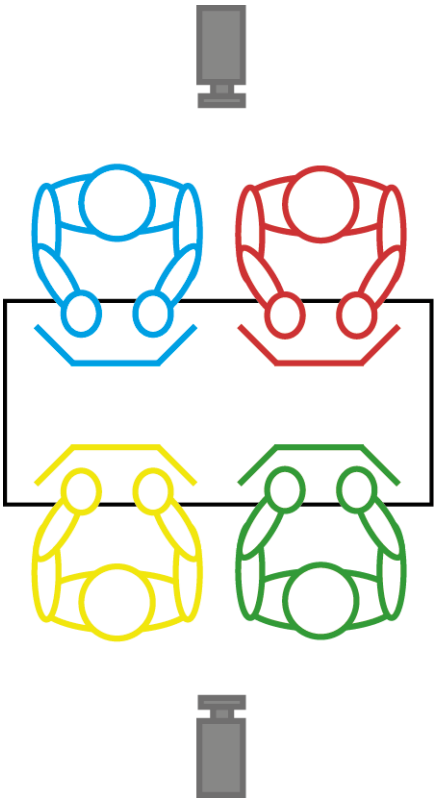


Figure 2.3. Player arrangement in the face-to-face context. Colour-coded screens ensured player privacy in terms of contributions, while allowing social interaction.

As in Experiment 1, participants completed five rounds of the game without punishment options to establish the procedures and ensure that all players understood the rules. After Round 3, they completed a short quiz to assess their understanding of the game and then received a standard reminder about game strategy. After the five practise rounds, the experimenter introduced the opportunity to punish free-riding players. Each player received four color-coded plastic punishment tokens. If a player decided to punish another, he/she placed one token into the contribution box, along with his/her own contribution. Upon receiving a contribution box containing a punishment token, the banker applied punishment to the player who had made the lowest contribution on that round. Players were aware that their

punishment would be applied to the lowest contributor on this current round. To indicate that they had been the lowest contributors, punished players received a black punishment token with their returns.

The same three punishment conditions as in Experiment 1 were active here. The experimenter was blind to which punishment condition would be active in the game until Round 5 when players received verbal instructions about the punishments. In the monetary punishment condition, punished players received the black punishment token in their contribution box along with a reduced return (as in Experiment 1). Because punishment tokens were returned privately, the identity of the punished player remained anonymous.

In the social punishment condition, punished players received the same return as the other group members; however, the experimenter placed the black punishment token into a pocket at the top of the punished player's screen. This meant that it was in full view of other players. In addition, the experimenter named the colour of the player who received the punishment (e.g., "The blue player has received a punishment."). Thus, although there was no financial penalty in this condition, the identity of the punished player was evident to all players. The third punishment condition included both social and monetary punishment, meaning that both punishments were applied. Participants played 10 punishment rounds, after which they were paid, debriefed and dismissed as in Experiment 1.

Video ratings

To obtain an estimate of the shared positive affect exchanged between each round, we asked an independent sample of participants (N=146; 96 females; Mean age=20.21, SD=2.76) to view and rate videos of the groups' interactions as they waited for round feedback. Each participant rated 15 videos (randomly assigned) and each video was rated 4 (60 videos) or 5 times (390 videos).

Participants rated each video according to a set of characteristics on a 5-point Likert scale (1=Not at all; 5=A great deal; see Appendix D for the full questionnaire). These included how smooth/coordinated the interaction seemed, how engaged the players appeared, how excited they seemed, how much shared laughter/smiling there was, how much they group talked, and the level of tension in the group (reverse scored). They also rated the overall mood of the group (1=Very negative; 5=Very positive). A factor analysis confirmed that these items loaded onto a single factor (item loadings>.52) and Cronbach's alpha analysis showed that participants rated the videos in a highly reliable fashion (alpha=.89). We therefore computed an average "group positivity" score by averaging the ratings for each video. These scores served as the interaction quality ratings for the analysis of the Experiment 2 data (below).

Results and Discussion

Investment behaviour

Participants' investments across trials appear in Figure 2.4a. As in Experiment 1, there were no differences in average contribution prior to the introduction of punishment, $F(2,120)=0.40$, $p=.673$, $\eta_p^2=.01$, although the threat of punishment did increase contributions. We also found that punishment type was important, $F(2,120)=9.00$, $p<.001$, $\eta_p^2=.13$. Here however, we found that punishments with a social component maintained high investment levels, whereas monetary punishments did not (p-values<.002). There was no difference in the average contributions of the two groups with social components to the punishment ($p>.99$).

Punishment descriptive statistics

Over half of all players opted to use most/all of their punishment tokens when available, with 20.2% using 3 tokens and 37.9% using all 4. This meant that 94% of Monetary trials, 74% of Social trials and 95% of combined trials included the punishment of a player. This indicated a marked reluctance in using social punishments; a difference that is reflected statistically- $F(2,120)=5.95$, $p=.003$, $\eta_p^2=.09$ (average number of punishments dispensed, per inidividual: Monetary=3.08 (SD=1.14); Social=2.05 (SD=1.52);

Combined=2.73 (SD=1.35)). Players in the social punishment only condition gave significantly fewer punishments than did those in the monetary punishment only condition (p=.003). The social versus combined conditions did not differ (p=.078), nor did the monetary versus combined conditions (p=.683). Of all the trials that contained a punishment, punishments were inflicted on the distributor (i.e., they were self-punishments) on 20.21% of monetary trials, 20.27% of social, and 24.21% of combined trials.

Cost of punishment

As in Experiment 1, the average increase in voluntary contributions on rounds in which players chose to punish, relative to their standard rate contribution (punishment phase trials in which a player is not including a punishment, nor did they receive one on the previous roun), did not depend on punishment condition, $F(2,85)=1.48$, p=.233, η_p^2=.03 (see Figure 2.4b). One sample t-tests showed that these values were significantly greater than zero (t-values>5.76, p-values<.001). Thus, regardless of group, participants increased their voluntary contributions to a similar degree when they chose to punish, in order to reduce the chance of receiving a punishment they had administered. We also found a significant main effect of punishment condition generally, $F(2,85)=15.62$, p<.001, η_p^2=.27, such that players in the monetary punishment condition changed their contributions less than players in the other conditions (p-values<.001). The social punishment conditions did not differ from one another (p>.99).

However, this was qualified by a significant punishment cost (giving, receiving) by punishment type (monetary, social, combined) interaction, $F(2,85)=5.00$, $p=.009$, $\eta_p^2=.11$. Post-hoc analyses showed that players in both social punishment conditions significantly increased their contributions above their standard rate after receiving punishments (one sample t-values>10.80, *p*-values<.001. However, players in the monetary punishment only condition were reluctant to increase their contributions (one-sample $t(26)=1.27$, $p=.215$). Thus, in this enriched social environment, monetary punishment alone did not enhance contribution behaviour, whereas social punishments did.

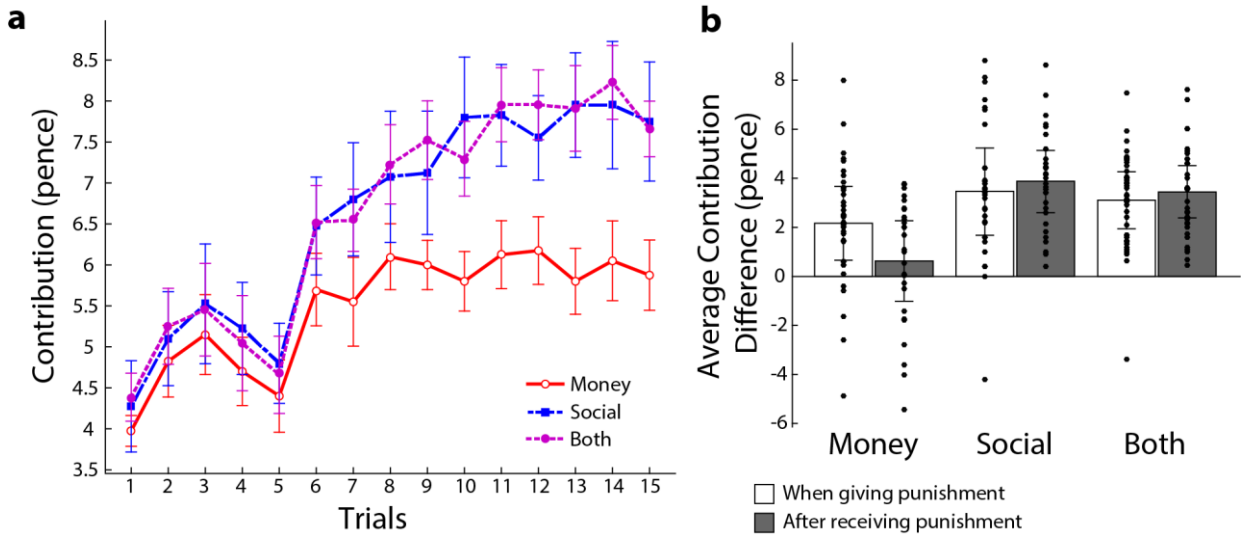


Figure 2.4. Experiment 2 Results. a) Contributions by punishment condition across trials. *NB*. This figure gives a descriptive representation of average individual contributions by round number, per condition, only. For statistical analysis, we analysed average contribution by punishment condition, collapsed across trials 6-15. b) Average increase in contributions when giving (white bars) and after receiving punishments (grey bars) by punishment condition. Individual data points are superimposed on the group averages. Error bars show the 95%CIs.

Sensitivity to inequality and positivity model

We also examined how players' sensitivity to inequality in contribution amounts shaped future investments, as in Experiment 1. However, in the context of this analysis, we additionally asked whether measures of group interaction quality (as rated from the interaction videos) on trial t predicted contributions on trial $t+1$. We estimated the following first order autoregressive model:

$$Y_{it+1} = C + \alpha(Y_{it} - \overline{Y}_i) + \beta(X_{jt} - Y_{it}) + \gamma(q_{jt}) + \varepsilon_{it}$$

In this model, C is the constant and models the intercept; α is the autoregressive coefficient, which models linear dependency in a time series. Y_{it} is participant i's contribution on trial t. $\overline{Y_i}$ is the average of all participant i's other contributions. Therefore, this term accounts for the extent to which a participant's contribution on each trial depends on his/her typical game strategy (e.g., participants who tend to make high contributions, regardless of others in the group). X_{jt} refers to the average contribution of the other players in group j on trial t, so the difference between X_{jt} and Y_{it} is the discrepancy between a player's contribution and the average contribution of other group members on that trial. β is the estimated regression coefficient for this term. This is modelled separately for each punishment condition. Finally, the term q_{jt} examines whether the group j participants' interaction quality between trials t and $t+1$, as rated by an independent sample of participants (see Experiment 2 – 'video rating' section), predicts their next contributions (γ describes the degree of this prediction). As above, we mean centered the raw data, in order to make meaningful comparisons amongst the estimated terms.

To determine the extent to which participants were sensitive to the discrepancy between their own investments and the average contributions of their group members, we examined the estimated model coefficients. These appear in Table 2.2. Interestingly, the data show that regardless of condition, participants were sensitive to discrepancies between their own contributions and the average of their fellow group members' contributions, after accounting for individual investment strategies. That is, they adjusted their next contribution based on contribution inequity in the present trial. Interestingly, the degree to which they did so was greater for the social and combined punishment conditions than it was for the monetary only punishment condition, as indicated by the 95%CIs on the estimates. The social and combined conditions did not differ from one another.

		Groups modelled together		
α (Autoregressive Component)	Regression Coefficient	0.79		
	95%CI: Lower Bound	0.72		
	95%CI: Upper Bound	0.85		
		Monetary Punishment	Social Punishment	Combined Punishment
β (Sensitivity to Inequality)	Regression Coefficient	0.17	0.35	0.37
	95%CI: Lower Bound	0.08	0.24	0.26
	95%CI: Upper Bound	0.25	0.46	0.47
γ (Interaction Positivity)	Regression Coefficient	0.19	0.31	0.33
	95%CI: Lower Bound	-0.07	0.01	0.10
	95%CI: Upper Bound	0.45	0.61	0.57

Table 2.2. Experiment 2 Regression Results (social setting). Estimated unstandardised model coefficients and 95%CIs.

Here, the model additionally included terms for the degree to which interaction positivity on a given trial predicted the next investment. A basic analysis of interaction positivity showed that although there was some variability across groups, $F(1,27)=3.35$, $p=.078$, $\eta_p^2=.11$, the differences across condition were not statistically significant, $F(1,27)=1.04$, $p=.368$, $\eta_p^2=.07$ (average positivity: Monetary=2.90 (SD=0.59); Social=3.19 (SD=0.56); Combined=3.24 (SD=0.68)). Nonetheless, interaction positivity on trial t for participants in the Social and Combined punishment conditions significantly predicted players' next contributions (regression coefficients greater than zero) but for the Monetary punishment condition it did not. However, the difference across the groups was not statistically significant, suggesting that positivity in social interactions generally drives up contributions (see Figure 2.5).

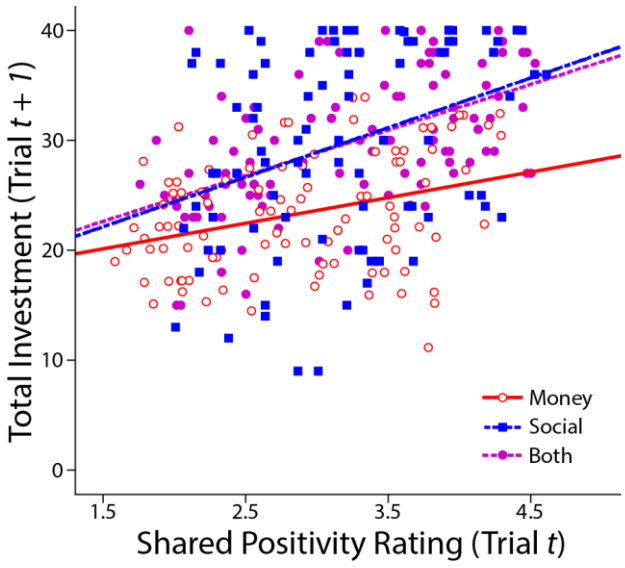


Figure 2.5. Scatter plot of interaction positivity on trial t and total group contribution on trial t + 1 across punishment condition.

Experiment 1 & 2 comparison

Average investment

To determine the effect of punishment type on average contribution and whether this differed depending on their social context, we compared investments across Punishment type (Money/Social/Both) and Interaction Condition (Anonymous/Face-to-face) in a 2-way between groups ANOVA. This analysis found main effects of both Interaction Condition, $F(1,238)=11.26$, $p=.001$, $\eta_p^2=.05$, and Punishment type, $F(1,238)=6.46$, $p=.002$, $\eta_p^2=.05$. Additionally, the interaction between the Interaction Condition and Punishment type was significant, $F(2,238)=9.32$, $p<.001$, $\eta_p^2=.07$. Bonferroni-corrected pairwise t-tests highlighted that this interaction was likely being driven by larger average investments when social and combined punishments were in effect during the face-to-face game, compared to the same punishment in the anonymous game (adjusted *p*-values <.001).

Furthermore, average investments in the Monetary punishment condition seemed to show the opposite trend in that contributions were marginally larger in the Anonymous game, compared to those in the Face-to-face game. However, the adjusted p-value exceeded the standard threshold for statistical significance ($p=.053$).

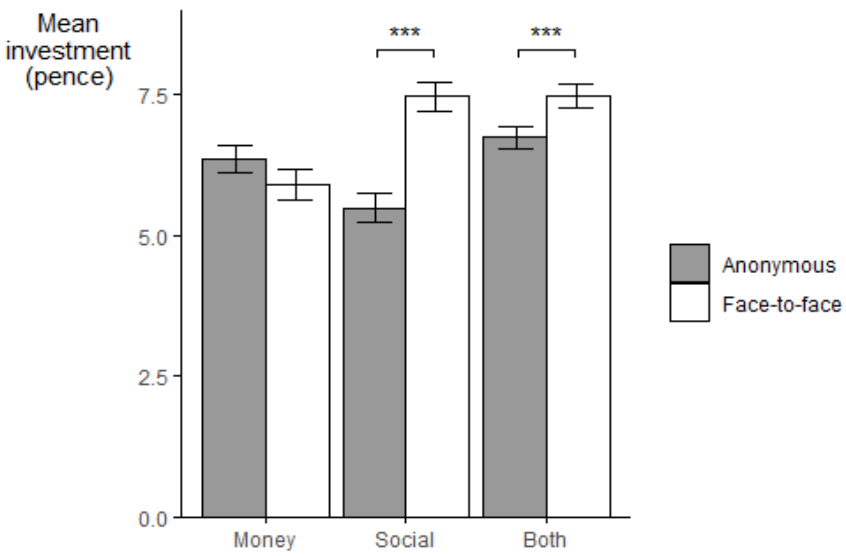


Figure 2.6. Average investment over interaction condition in the punishment game. Error bars show the 95%CI. ****p*-values <.001.

Together, these results suggest that the efficacy of punishment in raising voluntary contributions to the public good depended both on the type of punishment available and the setting in which participants were interacting. Specifically, social punishments help to raise

contributions in face-to-face environments, and monetary punishments are *marginally* more effective in anonymous game environments.

General Discussion

As predicted, findings from Experiment 1 show that in an anonymised interaction context, monetary but not social punishments served to maintain contributions to the public good over time. However, in the socially enriched, face-to-face context (Experiment 2) the pattern reversed, such that the threat of publically naming the lowest contributor encouraged contributions more so than than monetary punishments. Indeed, in the face-to-face condition, social punishments significantly enhanced the "cost" of receiving a punishment. Interestingly, in the anonymised condition, players raised their contributions after receiving punishment by similar amounts regardless of punishment type, even though these enhanced contributions were not sustained in the social punishment condition. These findings suggest that the cost of punishment depended on the interaction context. Specifically, punishments in the anonymised condition did not have the same sting as social punishments in the face-to-face version of the game. Additionally, in the face-to-face context of our study, participants interacted freely between rounds. When participants received social punishments in this context, group members' reactions were natural and unconstrained, often including friendly teasing, reciprocated smiles and laughter. This direct social feedback may account for the effectiveness of these punishments. Indeed, shared positivity (e.g., laughter, smiling) within a group's interaction on trial n predicted average contribution on trial n+1.

Broadly, results from both experiments align with literature suggesting that punishments are an effective moderator of cooperation, versus when the option for punishment is absent (for a review, see Balliet, Mulder, & Van Lange, 2011). Our work further demonstrates that the threat of punishment can influence contribution decisions, even in the absence of normative data (i.e. other players' contribution amounts), and extend prior findings

by showing that the social interaction context shapes punishment effectiveness. It is worth explicitly noting also that our results demonstrate how contribution behaviour changes under the *threat* of varying punishment types (as a function of social interaction context), as opposed to tracking its absolute effect on individual contribution decisions. In contrast to previous research in anonymous contexts (e.g., Masclet et al., 2003; Noussair & Tucker, 2005), we find that the threat of recceving a social punishment is singularly effective at maintaining contributions in social settings, even when contribution amounts are entirely unknown to players. This corroborates studies in laboratory and field settings suggesting that when both identity and reputation are at stake, behaviour becomes more cooperative (Andreoni & Petrie, 2004; Gächter & Fehr, 1999; Kraft-Todd et al., 2015; Milinski, Semmann, & Krambeck, 2002b), indicating that cooperative intentions can be signaled via social mechanisms, in addition to economic (i.e., contribution information).

Indeed, research suggests that social cues have the ability to alter economic decision-making. Smiles, for example, bias the estimation of reward probability, even when clearly unrelated to payoff (Averbeck & Duchaine, 2009) and cause people to overestimate the value of monetary rewards (Shore & Heerey, 2011). Smiling may also signal trustworthiness, thereby inducing cooperation in partners (Centorrino et al., 2015; Scharlemann et al., 2001). This idea is consistent with previous research suggesting that positivity is related to increased cooperation in public goods games (Rand, Kraft-Todd, & Gruber, 2015), as well as emotional expressivity more generally (Schug et al., 2010).

Our research therefore lends support to the idea that information present in the social milieu affects cooperative decision-making. We found that social punishments were costlier when others were able to observe punishment receipt. Evidence has suggested that the 'observability' of decision-making is an important factor in promoting cooperation in interaction partners (Jordan, Hoffman, Nowak, et al., 2016; Kraft-Todd et al., 2015; Yoeli et

al., 2013). As such, it may be the case that people use social cue information (e.g., smiles, emotional expressivity) to convey potential "cooperative intentions" to their interaction partners (Balliet et al., 2011; Jordan & Rand, forthcoming), or that the added value of this positive experience between decisions enhances subjective valuation of the return received (Averbeck & Duchaine, 2009; Shore & Heerey, 2011).

These ideas of observability closely relate to reputation and reciprocity incentives. That is, one's decisions to cooperate or defect in the real-world rarely occur without consequence. One's reputation, and thus the possibility of future reciprocity from social partners, is almost always at stake. Furthermore, reciprocity seems to explain the development of human cooperation (Nowak & Sigmund, 1998). That is, it may be worth paying the costs of cooperation now if this will ultimately be beneficial in the future. Nonetheless, this effect is contingent on one's good behaviour being communicable to others (Rand & Nowak, 2013). Moreover, situations that engage reputation concerns can enhance intuitive decisions to cooperate (Kraft-Todd et al., 2015).

Our experiment adds an important element to the standard public goods paradigm: the presence of naturalistic social interaction. This work therefore contributes to the body of literature that explores how social and affective information influence cooperative decisions. Specifically, we suggest that positivity may play a role here. However, more work is needed to explore the role of social information in cooperative decisions and how displays of specific social cues shape evaluations of social behaviour.

It is worth noting that we have not intentionally matched the 'strength' of punishments across social and monetary contexts here. Rather, the aim of the research presented here was to establish that the effectiveness of different types of punishment (social, monetary) strongly depend on the social context within which they occur. An additional point to note is that we did not directly assess participants' comprehension of the

punishment/reward mechanisms. So, we do not have any explicit insight into how well participants understood the consequences of including a reward/punishment, which would be helpful in decoding their decision-making process. In future studies, we will include a quiz specifically to assess understanding of punishment/reward at the end of the experiment.

Additionally, we assessed participant comprehension of the game incentives via a quiz question administered on round 3 (see Appendix C). Although we thought that this question was sufficient in providing an indication of whether participants in the group had a basic understanding of the game, it may be the case that some participants' game knowledge was not fully assessed. Thus, it is possible that there was variation in participant understanding that was not captured in the quiz results. In future research, we will assess comprehension with additional questions that more unambiguously indicate a subject's understanding of the game.

One final point to note is that we made some necessary alterations to the 'standard' game paradigm as implemented in economics literature. Specifically, we removed the "normative" cues used in most experimental versions of the public goods game that allow targeted punishments (e.g., Masclet et al., 2003; Noussair & Tucker, 2005; Rand, Dreber, Ellingsen, Fundenberg, & Nowak, 2009). Punishments in our game could not be targeted to particular players because our participants had no idea what their group members had contributed. Thus, punishment in our game version may not have provided participants with a strategic incentive to cooperate in the same way as do other games in the literature.

However, in spite of this change, it is interesting to note that our analysis did indeed find that participants were nonetheless responding to these incentives, and that their responses did differ depending on both the punishment type and social environment. This procedure provides insight into how humans may respond to the sorts of social consequences of free-riding (e.g., teasing, disapproval) that naturally occur in face-to-face settings. These

consequences disincentivise non-cooperation when recipients are known to each other and may serve to promote social cooperation in small groups.

Conclusion. In anonymous interaction conditions, social sanctions do not appear to maintain cooperation over time in the same way that monetary sanction do. However, in enriched social contexts, social punishments are especially costly, even under conditions in which no information regarding other players' contributions was present. We suspect that the immediacy of the social environment and threat to reputation present in this condition accounts for these results and suggest that when promoting cooperation, the best strategy is to match the type of sanction to the social context within which the cooperation should occur.

Chapter 3

The effect of social and monetary rewards on cooperation in the public goods game

Philippa J Beston

Erin A Heerey

Author contributions: P.J.B and E.A.H designed experiments, performed research and conceived of analyses; P.J.B. implemented 'investment behaviour' and 'cost of punishment' analyses. E.A.H implemented 'sensitivity/positivity' analyses. P.J.B wrote the paper. E.A.H proofread and suggested edits.

Abstract

The role of punishments in cooperative decision-making are often studied alongside that of reward. Research suggests that punishments are just as effective as rewards, however, they can also be destructive, with group earnings being healthier when rewards are available instead (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). Rewards are often monetary; however, a good reputation can also be rewarding. In this set of experiments, we compare Monetary (50% bonus to the highest contributor), Social (public naming of the highest contributor) and a combined type of rewards, using the same methodology as the previous chapter. Groups of participants played a Public Goods Game in either an anonymised or face-to-face environment. We found that in the anonymous game, the effect of rewards on cooperation did not differ by condition. Players raised their contributions in all conditions when distributing a reward. As the highest contributor was automatically rewarded, this appears to be active reward-seeking behaviour. In the face-to-face game, Monetary rewards were more effective at maintaining cooperation than Social. Similarly, players only increased contributions after giving a reward in the Monetary and Combined conditions. Furthermore, shared group positivity predicted contributions in these two conditions, but not in the Social condition. Taken together, it seems that participants were keen to earn rewards in monetary conditions, but tended to avoid social rewards in face-to-face contexts. Perhaps, the explicit reputational incentives crowded out cooperation, or our sample of primarily British subjects were reluctant to single themselves out for good behaviour.

Introduction

The effects of reward incentives on cooperation are often studied alongside those of punishment (for a meta-analytic review, see Balliet, Mulder, & Van Lange, 2011). The idea that cooperation deteriorates in the absence of either of these incentives is well supported (Fehr & Gächter, 2000a; Nowak & Sigmund, 1998b; Ostrom et al., 1992; Rand, Dreber, Ellingsen, Fundenberg, et al., 2009), as is the efficacy of costly punishment in reforming non-cooperative behaviour (Boyd & Richerson, 1992; Fehr & Rockenbach, 2004; Gintis, 2000; Gintis, Bowles, Boyd, & Fehr, 2003). Whilst costly punishment is effective, it can also be destructive. If used antisocially, punishments can backfire, serving to reduce levels of cooperation (Herrmann, Thöni, & Gächter, 2008), smother the potential for reciprocity between partners (Fehr & Rockenbach, 2003) and lower payoffs for the whole group (Dreber et al., 2008; Egas & Riedl, 2008; Szolnoki & Perc, 2010). However, group earnings are often healthier when players can reward others for cooperative behaviour, rather than punish non-cooperative behaviours (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009).

This would seem to suggest that rewards are more effective at maintaining cooperation than punishment. Indeed, early studies support this idea (Komorita & Barth, 1985), with recent studies finding that rewards reduce the detrimental retaliatory behaviour that punishment can provoke (Dreber et al., 2008; Nikiforakis, 2008). However, it appears that rewards are not able to sustain long term cooperation levels in the way that punishments can (Sefton et al., 2007). This may be because rewards do not target or alter the behaviour of non-cooperative players, whereas punishments do (Andreoni et al., 2003). Moreover, under conditions in which rewarding is costly, it is inefficient as it requires the continued sacrifice of resources to reward cooperators, whereas once defection has been stamped out, the costly punishment of non-cooperators is no longer necessary, meaning that punishers no longer need to sacrifice their own resources (Szolnoki & Perc, 2010).

Despite these negative aspects of reward use, players are prepared to reward other generous players, even at a cost to themselves (Almenberg et al., 2011). Moreover, individuals also seem to prefer incentive systems that dispense rewards rather than punishments (Sutter et al., 2010), are more willing to reward cooperators than to punish non-cooperators (Molenmaker et al., 2014) and are also more supportive of players who are rewarders than punishers (Kiyonari & Barclay, 2008). When given the choice, players even prefer to reward others with the most generous versions of rewards available (that have a cost-to-benefit ratio of 1:5, rather than 1:1); and these are also more effective at maintaining cooperation (Vyrastekova & Van Soest, 2008). Similarly, individuals seem to reward honesty more intensely than they punish deception, suggesting that this positivity has a stronger effect on behaviour (Wang, Galinsky, & Murnighan, 2009). Thus, although rewards appear to have a similar effect on levels of cooperation as punishment, compared to when either incentive is absent (Balliet, Mulder, & Van Lange, 2011), rewards may have a reduced psychological effect compared to punishments, meaning that individuals are more likely to reward than punish (Baumeister et al., 2001).

One important requirement for these effects is that interactions must be repeated, otherwise the threat of punishment or the promise of reward is effectively empty; the effects are short-lived and fail to foster cooperation (Walker & Halloran, 2004). Indeed, research also suggests that cooperation increases when dilemmas are iterated with consistent group members (Balliet et al., 2011), and that the opportunity for ‘targeted interactions’ (i.e., the ability to punish/reward specific players) is a key factor in cooperation as such opportunities introduce consequences for non/cooperative behaviour (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). These interactions also allow each player to develop a reputation over game rounds, which is an important factor in maintaining cooperation (Fu, Hauert, Nowak, & Wang, 2008; Sigmund, Hauert, Nowak, &

Wachter, 2001). The acquisition of a good reputation may even be a form of reward in itself, as known cooperators are often the beneficiaries of cooperation from others both within and outside of their own social groups (Semmann et al., 2005). Indeed, evidence from fMRI studies also suggests that the brain may process aspects of social reward, such as a good reputation, similarly to monetary rewards, suggesting that there is a common neural currency for rewards across these domains (Izuma, Saito, & Sadato, 2008).

Although research has suggested that gaining a reputation for being cooperative may be a form of reward in itself, there has been little research that formalises this idea. Here, we were interested in looking at the efficacy of rewards that convey reputational consequences (i.e., publically announcing the highest contributor on each round) and how this may differ from that of monetary rewards. In a similar vein to that of the previous chapter, we were also interested in extending these ideas to environments that allowed for realistic, face-to-face social interaction, compared to those where social partners were anonymous. Essentially, we predicted that results would mirror those from the previous chapter. We expected that in anonymous environments, 'reputational' rewards may not incentivise cooperation in the same way that they would in social contexts. Thus, we expected that in the anonymous version of the game, levels of cooperation may be lower when 'social' rewards were available compared to when economic rewards were available. We also predicted that results from the face-to-face version would suggest that people would be willing to cooperate more when they could gain a reputation for doing so, rather than when they can be materially rewarded for good behaviour (i.e., monetary rewards, as per Kraft-Todd, Yoeli, Bhanot, & Rand, 2015).

Experiment 1

In this experiment, we implemented a 15-round Public Goods game under anonymised conditions, similar to that described in Chapter 2, Experiment 1. Here

however, instead of offering players the opportunity to punish other players, it was possible to distribute and receive rewards. These rewards were not costly to administer, and were always awarded to the highest contributor of the round. As in the previous chapter, rewards and could be purely reputational (announcement of the identity of the highest contributor, henceforth labelled 'Social' rewards for simplicity), purely financial (a 50% monetary bonus: labelled 'Monetary' rewards here) or a Combination of the two. As above, we were interested in whether the efficacy of these rewards would differ depending on both reward type and the environment in which they were administered (i.e. when interactions were anonymised).

Method

Participants

One hundered twenty undergraduate participants completed a public goods game in exchange for partial course credit and a small monetary bonus. The sample consisted of a total of 128 participants (77 females; Mean age = 21.18, *SD* = 4.71); 10 groups of 4 players in the Monetary reward condition, and 11 groups each in the Social and Combined reward conditions. Participants provided written informed consent before participating and Bangor University's Ethics Committee approved all study procedures (likewise for Experiment 2 below). The sample size was determined in advance based on both budgetary limitations and the number of groups we estimated we would be able to recruit over the course of the year (likewise for Experiment 2).

Procedure

The procedure of this experiment is the same as that described in Chapter 2, Experiment 1. We adapted the methodology to accommodate using rewards rather than punishments; all other details remain the same. Participants attended the experiment in groups of four for an iterated 15-round public goods game, played online via networked computer terminals. Participants were distributed throughout a fully operational, busy

campus computer lab. Players were not aware of which other lab users were playing the game, nor their colour identity, but were aware they would play the same individuals throughout the game and that they would receive bonus money equal to their game earnings.

The same purpose-built website coordinated the game as was used in the punishment version of the game. Participants received a virtual endowment of 10 pence (0.10GBP) on each round, and then chose their contribution amount, by clicking the appropriate radio button (see Figure 2.1). They retained the rest in a private "bank". After the contribution phase, the database tallied the total fund, multiplied it by 1.6 and calculated each participant's return (always 25% of the total fund). Returns were rounded up to the nearest integer. The total group investment and the individual return were displayed at the end of the round. Participants played five practise rounds (no reward available). After Round 3, they completed a quiz assessing their understanding of the game and then received a standard reminder about the "rational" contribution strategy (Appendix B(i). See also Appendix B(iii) for comprehension analysis).

After five practise rounds, the website introduced the opportunity to reward other players. Players received four reward tokens to give to other players; these were not replaced if a player chose to distribute them all. Beginning on the contribution page for Round 6, if a player decided to reward another, he/she ticked a button on the screen, on the same screen as they had made their contribution decision. This reward was then anonymously applied to the player who had made the highest contribution on that round, randomly selecting amongst equally high contributors. A rewarded player received a black reward token icon and text feedback with return information (e.g., "On this round, YOU have received a reward!"). Only one player was rewarded on any one round.

There were three reward conditions: monetary only; social only; and monetary and social rewards combined. Reward type was a randomly assigned, between-groups

manipulation, so all members of a group experienced only one type of reward. Importantly, because the computer randomly assigned reward conditions, the experimenter was blind to which reward condition was active within a group.

In the monetary-only reward condition, rewarded players received a black reward token and text display on their feedback page, along with an increased return (e.g., "The GROUP return is [12] pence. YOUR return is [18] pence."). The increased return equalled the an additional 50% added to the individual's return. Importantly, only the rewarded player knew about the reward. Therefore, the identity of the rewarded player remained anonymous in this condition.

In the social reward condition, rewarded players received the same return as the other group members; however, the computer revealed the colour identity of the rewarded player, i.e., the highest contributor, to all players (e.g., "On this round, the RED player has received a reward!"). The third reward condition included both social and monetary reward, meaning that both reward types (publication of the rewarded player's colour along with an increased return) were applied.

As with the punishment version of the game, our players did not "pay" in order to apply a reward. The highest contributor was rewarded, which meant that if the player who chose to reward was *also* the highest contributor, he/she was the person who received the reward; likewise, for Experiment 2, below. Participants played 10 rounds of the game with reward options available. After this, the game ended and participants were debriefed, paid their game earnings and dismissed.

Data analysis

To determine the costs participants were willing to bear to administer rewards, we calculated the average contribution for each player for each trial in which that player administered a reward and compared that value to the player's average standard rate

contribution on trials in which the option to reward was available, but where the player had not chosen to give a reward, and had *not* just received one on the previous round. We also calculated how much players increased their contributions on the trial after receiving a reward (again, relative to their standard rate contribution), which provides a metric of the degree to which a reward is pleasant. If rewards are effective at enhancing contributions, a punished player's next contribution after reward receipt should be significantly higher than his/her typical standard rate contribution. We used a mixed-model design to examine the costs of rewards (changes in contributions relative to the average standard rate contribution; likewise, for Experiment 2). Reward Cost (for Giving versus Receiving a reward) served as the within-participants variable. Reward Condition (Monetary, Social, Combined) was the between-participants factor. Post-hoc comparisons are Bonferroni corrected throughout the report.

Results and Discussion

Investment behaviour

There were no differences in average individual investments across reward conditions during rounds 6-15, $F(2, 125)=0.44$, $p=.644$, $\eta_p^2=.01$. There were also no differences in contributions between reward conditions *before* rewards were introduced $F(2, 125)=0.16$, $p=.851$, $\eta_p^2=.01$. See Figure 3.1a for average player investments over trials 1-15. Note that in this analysis, we are not making any comparisons between the pre-reward (trials 1-5) and the reward (trials 6-15) phases. We have only analysed the pre-reward phase to compare average investment across reward groups. This was to ensure that groups were behaving similarly prior to the introduction of rewards – i.e., that there had been no administration error or experimenter bias that systematically affected the groups before the reward phase began.

Reward descriptive statistics

The average number of rewards distributed during trials where it was possible to reward another player did not appear to differ by reward condition (M_{money} = 3.28, *SD* = 1.09; M_{social} = 3.07, *SD* = 1.23; $M_{combined}$ = 3.05, *SD* = 1.18), which was confirmed statistically F(2, 125)=0.48, *p*=.618, η_p^2=.01. The majority of players opted to use most of their reward tokens, with 17.20% using 3 (of 4) tokens and 55.50% using all 4 tokens. Thus, nearly all trials included a reward - 95% of Monetary, 95.45% of Social, and 96.36% of Combined trials. Of all the trials where someone received a reward, players ended up rewarding themselves on 47.37% of Monetary trials, 42.86% of Social trials, and 38.68% of Combined trials.

Cost of rewards

There were no main effects of either cost, F(1, 57)=1.60, *p*=.212, η_p^2=.03, or reward condition, F(2, 57)= 2.04, *p*=.140, η_p^2=.07; nor was there an interaction between these variables, F(2, 57)= 0.19, *p*=.830, η_p^2=.01.

However, when comparing the average change in contribution per player to zero, i.e., an individual's standard rate contribution, it appears that in the Monetary condition, players put in significantly more than their standard rate after both giving and receiving reward (one-sample t-values ≥4.18, *p*-values≤.008). In the Social and Combined conditions, players contributed more after giving (one-sample t-values ≥3.01, *p*-values≤.013), but not receiving a reward (one-sample t-values ≤1.96, *p*-values ≥.013), see

Figure 3.1b.

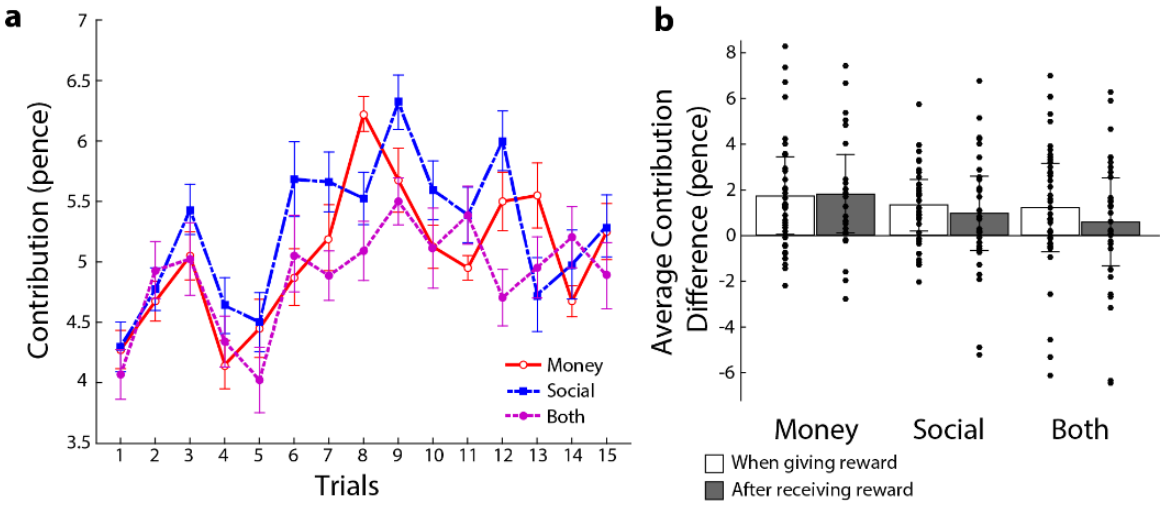


Figure 3.1. Experiment 1 results. a) Contributions by reward condition over trials 1-15. b) Average increase in contributions after players gave (white) or received (grey) a reward, by reward condition. Individual data points are superimposed on the group averages. Error bars show the 95%CIs.

Sensitivity to inequality model

As in Chapter 2, Experiment 1, we were interested to learn whether participants become aware of inequality between group member contributions even in the absence of descriptive information (as evidence shows that they are sensitive to such discrepancies when normative information (i.e. other players' contribution information) *is* present; Chaudhuri, 2011; Fehr & Gächter, 2000). Specifically, we used a model of the data to examine how the discrepancy between one's own contribution and the average of group member contributions on trial t predicted contributions on trial $t + 1$, whilst accounting for an individual's typical contribution strategy (autoregressive component). We call this sensitivity to inequality. As participants become aware of the fact that their contributions approach 25% of the total pool, they should use this information to infer other player's likely contribution strategy, thereby adjusting their own (see page 30 for more details about model parameters). Table 3.1 displays results from this regression model, where we examine the strength of this estimated term. Results suggest that only in the Combined reward condition did participants become aware of the discrepancy between their

contributions and the average group contribution and use this information to determine their next contribution. The regression coefficients for the Monetary and Social conditions were not different from zero, suggesting that in these other conditions, they did not.

		Groups modelled together		
α (Autoregressive Component)	Regression Coefficient	0.57		
	95%CI: Lower Bound	0.49		
	95%CI: Upper Bound	0.64		
		Monetary Reward	Social Reward	Combined Reward
β (Sensitivity to Inequality)	Regression Coefficient	-0.04	0.08	0.15
	95%CI: Lower Bound	-0.13	-0.01	0.06
	95%CI: Upper Bound	0.05	0.18	0.24

Table 3.1. Experiment 1 Regression Results (anonymous setting). Estimated unstandardised model coefficients and 95%CIs.

Experiment 2

This experiment follows a similar methodology to that of Chapter 2, Experiment 2. Here, we were interested in extending the methods of the previous experiment to a setting that involved face-to-face social interactions. Players could choose to reward the highest contributor on each round in one of three different ways: to incur a Monetary reward, a Reputational or a Combined version of both of these reward types. These reward conditions were the same as those available in the previous experiment, but here players could see and interact with others during their decision to cooperate.

Method

Participants

We recruited 120 participants (76 females; Mean age = 20.93 *SD* = 4.48) from an undergraduate student research pool. Participants received partial course credit and a small monetary bonus based on their earnings in the game.

Procedure

This version of the game used the same basic procedure to that in Chapter 1, Experiment 2, with adaptations for the use of rewards. Here participants played a 15-round public goods game face-to-face, using real pennies as monetary tokens. Participants were

seated around a table; clearly visible to all group members. All interactions were video recorded, but cameras were positioned so that investment behaviour remained private.

Players received a colour-identity that they used throughout all 15 rounds. They sat behind colour-coded screens to conceal their contribution behaviour from the rest of the group. On each round, participants chose the amount that they wished to contribute to the group resource from their endowment of 10 pennies. They made their contributions by placing pennies into opaque, color-coded boxes, passing these to the experimenter, who then passed them to a second experimenter playing the role of the "banker" in an adjacent room. This procedure ensured that players' contribution amounts and strategies were hidden from the experimenter who interacted with them and that the banker was blind to player identity.

The banker calculated the total fund and computed the group's return (rounding up to the nearest whole penny, if necessary), placing return amounts into the players' contribution boxes, and then returned these to the experimenter. The experimenter announced the total contribution and individual return on the round (e.g., "The total was 15 pence and your return is 6 pence."). Players deposited their return and any endowment remainder into their individual banks and a new round then began. Players were unaware of the number of rounds they would complete.

Participants could interact freely whilst the banker counted contributions. As above, however, explicit negotiation about game strategy/contributions was forbidden. Measures of interaction mood and quality were rated from videos of these interactions.

As in Experiment 1, participants completed five rounds of the game without reward options, completing the quiz at the end of Round 3, and then receiving a standard reminder about game strategy. After the five practise rounds, the experimenter introduced the opportunity to reward other players. Each player received four coloured-coded plastic reward tokens. If a player decided to reward another, he/she placed one token into the

contribution box, along with his/her own contribution. Upon receiving a contribution box containing a reward token, the banker applied reward to the player who had made the highest contribution on that round. Players were aware that the reward would be applied to the highest contributor on this current round. To indicate that they had been the highest contributors, rewarded players received a gold reward token with their returns.

The same three reward conditions as in Experiment 1 were active here. Importantly, the experimenter was blind to which reward condition would be active in the game until Round 5 when players received verbal instructions about the rewards. In the monetary reward condition, rewarded players received the gold reward token in their contribution box along with an increased return (as in Experiment 1). Because reward tokens were returned privately, the identity of the rewarded player remained anonymous.

In the social reward condition, rewarded players received the same return as the other group members; however, the experimenter placed the gold reward token into a pocket at the top of the rewarded player's screen. This meant that it was in full view of other players. The experimenter also named the colour of the player who received the reward (e.g., "The blue player receives a reward."). Thus, the identity of the rewarded player was evident to all players. The third reward condition included both social and monetary reward, meaning that both reward types were applied. Participants played 10 reward rounds, after which they were paid, debriefed and dismissed as above.

Video ratings

To obtain an estimate of the shared positive affect exchanged between each round, we asked an independent sample of participants (N = 174) to view and rate videos of the groups' interactions as they waited for round feedback. Each (randomly assigned) video was rated 5 or 6 times.

As in Chapter 1, participants rated videos on a 5-point Likert scale on several characteristics such as smooth/coordinated the interaction seemed, how engaged the players appeared, how excited they seemed, how much shared laughter/smiling there was, how much they group talked, and the level of tension in the group (reverse scored) as well as the overall mood of the group (see Appendix D for the full questionnaire). As with Chapter 1, Experiment 2, we computed an average "group positivity" score by averaging the ratings for each video.

Results and Discussion

Investment behaviour

When looking at investments during rounds 6-15 (i.e. when rewards were available), we found significant differences between reward groups, $F(2, 117)=8.13$, $p<.001$, $\eta_p^2=.12$. Investments in the social reward condition were significantly lower than those in the monetary reward condition ($p=.001$) and also lower than those in the combined condition ($p=.008$). There was no difference between the monetary and combined reward conditions ($p=1$); see Figure 3.2a for average player investments by trial. There were no differences in contribution levels across reward conditions before rewards were available $F(2, 117)=0.70$, $p=.499$, $\eta_p^2=.01$.

Reward descriptive stats

The average number of rewards distributed per player appeared to be similar across reward condition ($M_{money} = 3.10$, $SD = 1.15$; $M_{social} = 2.98$, $SD = 1.03$; $M_{combined} = 3.15$, $SD = 1.10$) and did not differ statistically, $F(2, 117)=0.27$, $p=.762$, $\eta_p^2=.01$. When rewards became available, players opted to use the majority of them, with 33.3% of players using 3 tokens, and 44.2% using all 4 tokens. As with the previous experiment, nearly all trials included a reward (97% of Monetary, 95% Social, 95% Combined trials). Players both included and received a reward (i.e. a self-reward) on 50.52% of Monetary trials, 38.95%

of Social, and 60% Combined of trials. These statistics would indicate that players sought rewards less in the Social reward condition.

Cost of rewards

When analysing the cost of rewards (to Give and Receive) by reward condition (Money, Social, Combined), we found that there was a main effect of Cost type $F(1, 71)= 65.88$, $p<.001$, $\eta_p^2=.48$. It appears that regardless of reward condition, players contributed more than their standard rate when giving compared to receiving a reward, suggesting active reward seeking behaviour. There was also a significant main effect of reward condition, $F(2, 71)= 14.22$, $p<.001$, $\eta_p^2=.29$, with post-hoc tests suggesting that there was a significant decrease in contributions in the social relative to the monetary and combined reward conditions, (p-values<.001). This analysis also showed an interaction between reward cost and reward condition $F(2, 71)= 3.36$, $p=.040$, $\eta_p^2=.09$, which appears to be driven by the difference in contributions between giving and after receiving a reward across the conditions, see Figure 3.2b.

In conditions where monetary rewards were available (Monetary and Combined conditions), on average, players contributed significantly more when giving a reward (one-

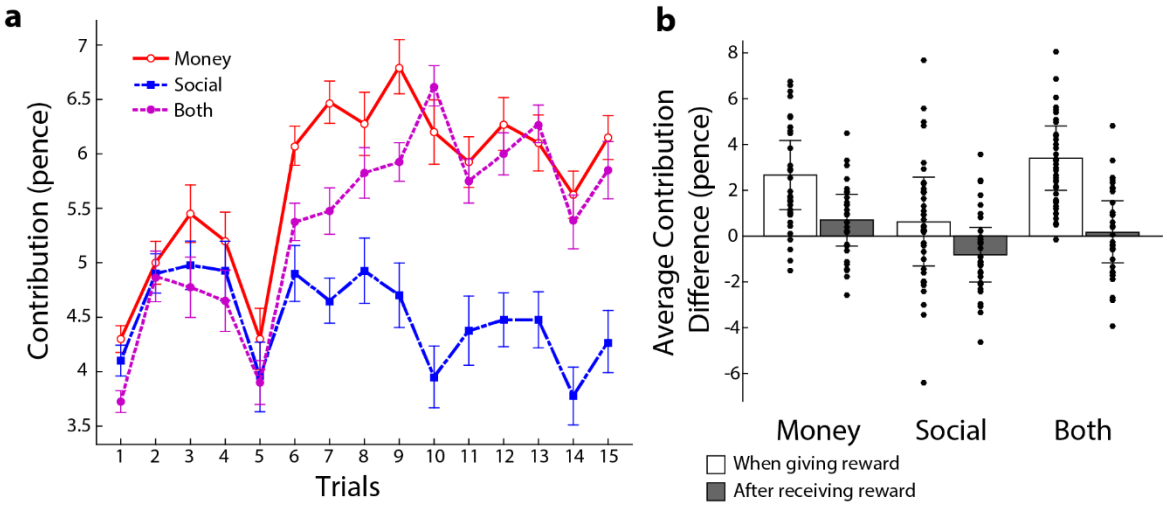


Figure 3.2. Experiment 2 results. a) Contributions by reward condition over trials 1-15. b) Average increase in contributions after players gave (white) or received (grey) a reward, by reward condition. Individual data points are superimposed on the group averages. Error bars show the 95%CIs.

sample t-values>5.32, *p*-values <.001), suggesting payoff driven reward seeking behaviour. Players did not contribute more after receiving a reward in these conditions (one-sample t-values≤1.80, *p*-values ≥.106). Players did not contribute more in the Social condition when giving ($t(9)=1.06$, $p=.315$) or receiving a reward ($t(9)=-1.87$, $p=.094$), suggesting that rewards had little effect on economic behaviour in this condition.

Sensitivity to inequality, and positivity model

As in Experiment 1 we used a regression model to examine sensitivity to inequality. Here we also added an extra term to the model that examines whether group interaction quality between trials t and $t+1$, as rated by an independent sample of participants (see 'video rating' section above), predicts subsequent contributions (see Chapter 1, Experiment 2 for more details about model parameters). Table 3.2 displays results from this model.

		Groups modelled together		
α (Autoregressive Component)	Regression Coefficient	0.67		
	95%CI: Lower Bound	0.59		
	95%CI: Upper Bound	0.75		
		Monetary Punishment	Social Punishment	Combined Punishment
β (Sensitivity to Inequality)	Regression Coefficient	0.21	0.21	0.35
	95%CI: Lower Bound	0.12	0.10	0.26
	95%CI: Upper Bound	0.31	0.31	0.44
γ (Interaction Positivity)	Regression Coefficient	0.42	-0.04	0.62
	95%CI: Lower Bound	0.15	-0.34	0.24
	95%CI: Upper Bound	0.68	0.26	0.99

Table 3.2. Experiment 2 Regression Results (face-to-face setting). Estimated unstandardized model coefficients and 95%CIs.

In terms of sensitivity to inequality, results from this model suggest that, regardless of reward condition, players adjusted their next contribution based on inequity in the present trial (after accounting for individual investment strategies). Interestingly, although levels of positivity did not differ between reward condition, $F(2, 29)=0.49$, $p=.617$, $\eta_p^2=.03$; positivity on a given trial did predict investment on the next in the Monetary and

Combined reward conditions (regression coefficients greater than zero). However, this was not the case for the social reward condition, suggesting that regardless of interaction positivity, players were reluctant to increase their contributions on the next round even when experiencing positive interactions (see Figure 3.3). Thus, in this experiment it appears likely that without the financial incentive of receiving a reward, participants were generally unwilling to contribute extra money to gain a reputation as a generous player.

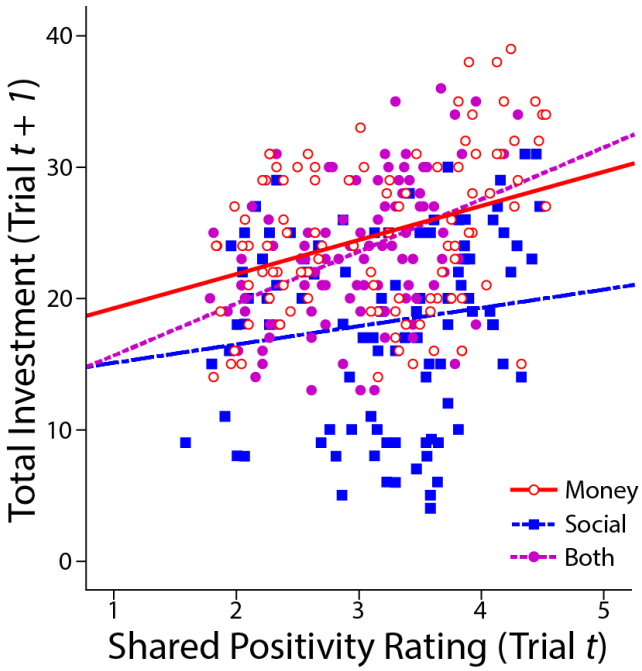


Figure 3.3. Scatter plot of interaction positivity on trial t and total group contribution on trial t + 1 across punishment condition.

Experiment 1 & 2 comparison

Average investment

To assess the effect of reward type and the social context in which it was administered on contribution levels, we conducted a 2-way between groups ANOVA[8]. The Interaction Condition (Face-to-face, Anonymous) by Reward type (Money, Social, Both) interaction term was significant - $F(2,242)=5.57$, $p=.004$, $\eta_p^2=.04$. Follow up pairwise t-tests further demonstrated that average participant investment was larger in the Face-to-face setting with Monetary/Combined versions of Reward (adjusted *p*-values<.001). However, in the Social reward condition, average contributions were larger in the Anonymous version of the game, compared to the face-to-face game (Bonferroni adjusted *p*<.001). Results from this analysis suggest that Social rewards were successful in increasing contributions, but only if they were not adminstered in a Face-to-face interaction context.

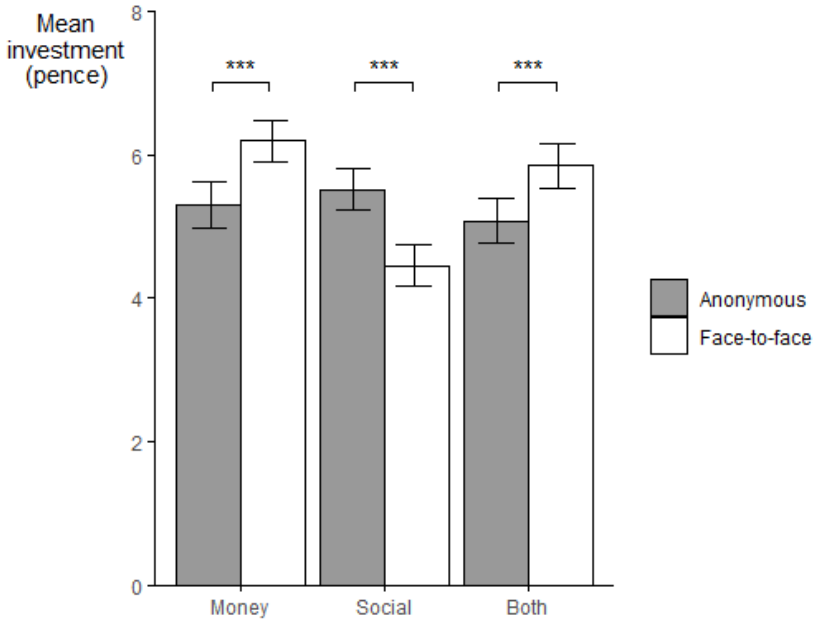


Figure 3.4. Average investment over interaction condition in the reward game. ****p*-values <.001. Error bars show 95%CI

General discussion

In these experiments, we were interested in comparing the effects of different types of rewards on contribution behaviour. Social rewards aimed to engage reputational

[8] This analysis is essentially a follow up to a 3-way ANOVA documented in Appendix E, which also included experiment type (reward/punishment) as an additional term in the model.

mechanisms, whereas Monetary rewards were to engage players' economic interests. We allowed players to interact in two interaction settings; either face-to-face or anonymous. Results from our anonymous Public Goods Game showed that reward type does not differentially affect investment behaviour; player contributions do not differ by reward condition. Here, players also seemed to raise their voluntary contributions to the public good after choosing to distribute a reward in all conditions, suggesting that they were willing to pay a cost to reward the highest contributor (which could have been themselves). However, players did not tend to contribute more after receiving a reward, suggesting that the receipt of a reward did not incentivise future contributions (except in the Monetary reward condition).

During the face-to-face version of the game, average contributions to the public good were significantly higher in the Monetary and Combined conditions, compared to Social. Players in these two conditions also contributed more when choosing to distribute a reward, suggesting reward seeking behaviour. In these two conditions, group positivity in social interactions predicted contribution behaviour, suggesting that these social interactions did indeed play a role in cooperative decision-making. Interestingly, we also found that in the Social reward condition, players experienced the same amount of positivity as in other conditions, but this failed to predict contribution behaviour, suggesting that even though players were getting on well with their interaction partners, they were still reluctant to increase their contributions. Similarly, it seemed that players in this condition did not contribute significantly more after giving or receiving a reward.

We predicted that results in this experiment would essentially replicate those of the previous chapter. We expected that in the anonymous game, rewards that engaged reputational incentives would not be especially effective, because of the limited opportunity to build a meaningful reputation as a cooperator here. We thus expected that participants

may be more economically driven in this environment and opt to contribute more when there was a financial incentive for doing so. In an environment where social partners can see and interact with each other, however, we expected that Social rewards would encourage higher contributions than in the monetary condition, due to enhanced reputational incentives, as per Kraft-Todd et al., (2015). Contrary to our hypothesis, we found that in the anonymous setting, cooperation levels looked similar across all reward conditions and that participants actively pursued rewards, suggesting that all reward types were equally effective in maintaining cooperation over time. In the face-to-face setting, we found the opposite effect to our prediction; monetary incentives were much more influential on contribution behaviour than social, and participants were only willing to pay to distribute monetary rewards.

These results seem to suggest that players tended to want to avoid receiving Social rewards in the face-to-face setting. One explanation for this result may have been that our British participants were reluctant to appear to actively seek these reputational rewards to avoid being singled out in front of other group members. Another explanation might be that the idea of gaining a reputational reward in this environment is 'crowding out' cooperation. For example, obvious monetary incentives negatively interact with reputational concerns to crowd out cooperation in public contexts as they may actually signal selflessness for selfish reasons (Ariely et al., 2009). This may also be the case when reputational incentives are overly explicit, as they obscure the validity of the signal that they are sending (Bénabou & Tirole, 2006). So, although engaging reputation mechanisms are important for the existence and maintenance of cooperation (Sigmund et al., 2001), it is also important that reputational incentives are not so overt that they clearly signal a selfish motivation for cooperation (Kraft-Todd et al., 2015).

Previous research about rewarding behaviour has suggested that players are willing to pay a cost to reward cooperative players (Almenberg et al., 2011). Both experiments

presented here do also seem to suggest that participants are generally willing to pay to distribute monetary rewards, especially when they know that increases in contributions enhance the chances that they will receive the reward themselves, thereby earning back their additional contributions. However, it is important to note here that rewards were automatically distributed to the highest contributor, so although participants raised their contributions when choosing to distribute a reward, it is likely there was a self-regarding, and economic motive (in the Monetary reward conditions), behind this rewarding behaviour.

To our knowledge, this is the first reward study in which participants were able to interact freely, in a face-to-face context, during cooperative decision-making. Some previous punishment studies feature some aspects of social interaction, but these were often overly structured interactions (e.g., Gächter & Fehr, 1999) or restricted to occur before a cooperation game (Gaertig et al., 2012), rather than occuring concurrently. Furthermore, much of this research also specifically examines the communication of *intentions* for cooperation (e.g., Arechar et al., 2017; Balliet, 2009; Isaac & Walker, 1988), these interactions being qualitatively different than exchanges that do not involve explicit bargaining (Putnam & Jones, 1982). One novel aspect of our research is that interactions in the face-to-face context were entirely unprompted; players could interact as much, or as little, as they wanted. There was also the opportunity for these interactions to occur throughout the whole experiment, which had the advantage that we could independently rate these social interactions on a round-by-round basis and model their effect on cooperation. Here we looked at the role of positivity especially, however, it may be interesting for future research to look at how negative affect influences cooperation.

One limitation of our research was that there was no interaction possible between players in the anonymous game, therefore we do not have a comparison measure of interaction quality in this setting. Future research could look at how the quality of web-based

interactions may or may not influence cooperation, especially as these kinds of interactions are becoming more common-place in an interconnected world.

Of additional note is that when running a post hoc power analysis, based on the relatively small achieved effect sizes of the 2-way analysis of average investment, the suggested sample size came out to 606 independent groups (101 per condition or 2424 individual participants). Our total N for this experiment was in fact 248 (approximately 10 groups of 4 players in each condition, as above). A sample of 606 participants was simply not feasible in the context of this design, given the relatively small participant pool available to us, the extensive costs in terms of testing time, and a lack of external funding for this PhD project. However, we had no way of knowing, *a priori*, that we would see such small effects in this context – instead we estimated an anticipated effect that was similar in size to the punishment effect and nowhere near as small as that we actually obtained. We therefore conclude that although our sample was insufficient to detect an effect in this context, we would simply have been unable to collect the type of sample size required to detect these small effects.

Conclusion

The research presented here looked at the role of reward in cooperative decision-making. Broadly, we found that regardless of interaction setting, in conditions where monetary rewards were available, players were willing to pay a cost in order to distribute rewards, perhaps in order to earn back this monetary bonus. We also found that in contexts where players can interact face-to-face with other group members, social/reputational rewards do not sustain cooperation as well as monetary rewards do. Players in the Social reward condition experienced a similar amount of group positivity in their interactions, however, this positivity did not predict contributions in this setting in the way that it did in the Monetary and Combined conditions. Based on these findings, we support the use of

monetary rewards for cooperative behaviour, regardless of interaction setting, but we advise against the use of social rewards in face-to-face environments as they appear to crowd out cooperation, and may even be aversive.

Chapter 4

Social feedback interferes with implicit rule learning: Evidence from event-related brain potentials.

Philippa J Beston

Cécile Barbet

Erin A Heerey

Guillaume Thierry

A version of this appears in published form as: Beston, P. J., Barbet, C., Heerey, E. A., & Thierry, G. (2018). Social feedback interferes with implicit rule learning: Evidence from event-related brain potentials. Cognitive, Affective, & Behavioral Neuroscience, 1-11. https://doi.org/10.3758/s13415-018-0635-z

Author contributions: P.J.B conceived experiment idea, P.J.B and G.T designed experiment, with input from E.A.H. P.J.B programmed experiment, with input from C.B and G.T. P.J.B acquired data. P.J.B analysed data with input from C.B. P.J.B wrote the paper, with input from G.T, and all co-authors proofread and provided suggestions for edits.

Abstract

The human brain can learn contingencies built into stimulus sequences unconsciously. The quality of such implicit learning has been connected to stimulus social relevance, but results so far are inconsistent. Here, we engaged participants in an implicit-intentional learning task in which they learned to discriminate between legal and illegal card triads on the sole basis of feedback provided within a staircase procedure. Half of the participants received feedback from pictures of faces with a happy or sad expression (social group) and the other half based on traffic light icons (symbolic group). We hypothesised that feedback from faces would have a greater impact on learning than that from traffic lights. Although performance during learning did not differ between groups, the feedback error-related negativity (fERN) was delayed by ~20 ms for social relative to symbolic feedback; and the P3b modulation elicited by infrequent legal card triads within a stream of illegal ones during the test phase was significantly larger in the symbolic than the social feedback group. Furthermore, the P3b mean amplitude recorded at test negatively correlated with the latency of the fERN recorded during learning. These results counterintuitively suggest that, relative to symbolic feedback, socially salient feedback interferes with implicit learning.

Introduction

Humans can learn contingencies about their environment without conscious awareness. Such phenomenon is classically observed in the case of statistical learning, when dependencies between linguistic stimuli, for instance, are extracted by the brain without the participants' intention to acquire them (Saffran, Aslin, & Newport, 1996). Such form of spontaneous and unconscious learning is observed across a variety of perceptual domains. In the auditory domain, for example, studies have shown that infants implicitly use language patterns to rapidly segment words from speech streams (Aslin, Saffran, & Newport, 1998; Saffran, 2003), and this phenomenon extends even beyond linguistic stimuli (Saffran, Johnson, Aslin, & Newport, 1999). Statistical learning occurs spontaneously (Fiser & Aslin, 2001), rapidly (Turk-Browne, Scholl, Chun, & Johnson, 2009), and without the need for explicit instruction (Fiser & Aslin, 2002). This has led to the proposal that statistical learning results in the formation of implicit knowledge (Fiser & Aslin, 2002; Perruchet & Pacton, 2006; Reber, 1967; Turk-Browne, Jungé, & Scholl, 2005).

Implicit learning refers to the process of learning the underlying rule of a system (e.g., an artificial grammar) solely based on exposure to stimulus contingencies and probabilities (Reber, 1967). Just like statistical learning, implicit learning is thought to be unconscious, meaning that participants are unable to verbalise a rule that they have acquired (Reber, 1989), and are not aware that they have learnt something (Cleeremans, Destrebecqz, & Boyer, 1998; Dienes, Altmann, Kwan, & Goode, 1995; Seger, 1994). For example, evidence from the serial reaction time task suggests that people identify previously viewed light sequences more quickly than novel sequences (Chun & Jiang, 1998; Willingham, Nissen, & Bullemer, 1989).

One limitation of much work within this literature is that the nature and quality of learning is measured by participant performance or metacognitive evaluations after learning. In the problem-solving domain, for example, implicit memory of a puzzle improves problem solving on a subsequent task, even when participants are given a concurrent task to exert

strain onto the working memory system (Reber, 1989; Reber & Kotovsky, 1997). The results of such studies appear to be contingent upon the type of task used to determine whether learning was implicit (Shanks & Channon, 2002; Wilkinson & Shanks, 2004). Thus, the extent to which the process is truly unconscious remains debatable. Still, when participants perform above chance after training, although they believe that they are merely guessing the answers, one may presume that the learning was mostly unconscious and that their knowledge is implicit (Dienes et al., 1995). Shanks and St John (1994) have questioned how much post-learning tests (e.g., asking participants to verbalise a rule that they have acquired) tell us about the nature of the learning process. More specifically, they enquired whether tests of performance and awareness are sensitive enough to measure the acquired knowledge that has become conscious and whether knowledge awareness can really be assessed before the nature of the knowledge itself has been determined. It seems that classic implicit learning tasks lack precision regarding the nature of what participants learn when the conclusions are solely drawn from performance indices, e.g., reaction time (Eimer et al., 1996) or post-learning verbalisations (Shanks & St John, 1994).

One way to obtain unbiased evidence of implicit learning is to measure spontaneous brain activity modulations elicited by learned contingencies. Event-related potentials (ERPs), a method derived from electroencephalography, are averaged recordings of brain activity measured at the surface of the scalp elicited by series of repeated stimuli. ERPs can be recorded independently of performance indices and index unconscious information processing in the absence of any behaviourally measurable effect (Thierry & Wu, 2007; Wu & Thierry, 2010). Baldwin and Kutas (1997), for instance, showed that participants engaged in an artificial grammar learning task (without any explicit instruction regarding underlying rules) produced P300 ERP responses of larger amplitude for correct grammatical forms than incorrect ones. This result shows that participants developed expectancies about the

sequences they viewed and were able to detect rule violations, even though they seemed unable to consciously access this information at debriefing (see also, van Zuijen, Simoens, Paavilainen, Näätänen, & Tervaniemi, 2006), Similar effects have also even been shown in cases where rule learning was not embedded within the experiment but rather occurred from natural exposure to language. For example, Vaughan-Evans et al., (2016) recently showed that the brains of individuals with no recorded or overt knowledge of an ancient form of Welsh poetry (Cynghanedd) successfully identified correct forms from sentences violating composition rules, despite being unable to detect the correct forms in overt judgement tasks. Presumably, these participants learned the rules of Cynghanedd implicitly, through natural language exposure and required no conscious knowledge.

One important dimension of the human learning environment that seems to have been neglected so far in the implicit learning literature is the social quality of the information people learn, even though it is reasonable to assume that feedback during learning would vary in efficiency depending on its social significance. Information from and about other humans is abundant in the environment, and even the mere presence of others has long been suggested to facilitate performance on simple tasks (Bond & Titus, 1983; Zajonc, 1965; Zajonc, Heingartner, & Herman, 1969). More recent research has also suggested that reliable social cues allow others to implicitly predict their behaviour, e.g., in a game of rock-paper-scissors (Heerey & Velani, 2010). Social cues such as smiles and frowns can also aid performance during associative learning as compared to non-social ‘traffic light’ feedback) (i.e., 'symbolic' feedback; Hurlemann et al., 2010). These findings suggest that socially relevant information is processed by the same associative system that underlies other types of reward-based learning (Behrens, Hunt, Woolrich, & Rushworth, 2008). However, during cognitively demanding tasks, participants avert gazing at faces, and the frequency of this avoidance relates to task difficulty (Doherty-Sneddon, Bruce, Bonner, Longbotham, &

Doyle, 2002; Glenberg, Schroeder, & Robertson, 1998). Thus, social information appears to add a cognitive load during difficult tasks and participants spontaneously resort to gaze averting in order to reduce this load (Doherty-Sneddon & Phelps, 2005).

Nevertheless, there is a distinction between learning that occurs within a social context (Glenberg et al., 1998; Heerey & Velani, 2010) and learning that results in a socially charged outcome, e.g., when socially relevant information conveys feedback about performance (Turnbull, Bowman, Shanker, & Davies, 2014). In the latter case, there is some indication of a facilitatory effect on performance in associative tasks (Hurlemann et al., 2010), however, it is unknown how performance is affected in tasks that require implicit contingency learning.

Here, we presented participants with triads of cards featuring coloured shapes, varying in 4 possible ways (shape type, colour, number of shapes, and filling) and asked them to indicate which triads were 'legal' combinations and which were 'illegal', according to a rule that was never described. Participants were thus engaged in an implicit-intentional learning task, in which they were instructed to proceed on a trial-and-error basis. We labelled this context as intentional because participants were aware of the need to extract some kind of rule, even though they did not know what this rule was. This task context notably differs from the incidental context that usually applies in classical implicit learning. They received feedback on every trial, completing the learning phase only when they had met a pre-determined learning criterion applied via a staircase procedure (described in the Procedure section). Participants received feedback with either faces or a traffic light display. Given their high social-relevance and the fact they have been shown to increase performance in associative tasks (Hurlemann et al., 2010; see also Mihov et al., 2010), we hypothesised that faces as feedback would boost performance during learning and result in higher accuracy in a subsequent test phase. Crucially, in order to collect an objective and spontaneous marker of

learning, we recorded ERPs throughout the two phases of the experiment and monitored: (a) the participants' physiological reaction to feedback (indexed by the feedback error-related negativity, fERN) during the learning phase, and (b) their spontaneous response to infrequent legal card combinations, presented amongst frequent illegal ones (as indexed by the P3b modulation elicited by infrequent stimuli in an oddball paradigm) during the test phase. Consistent with our hypothesis, we expected that face stimuli would elicit greater fERN amplitudes during learning, and thus lead to greater mean P3b amplitudes at test.

Method

Participants

Fifty-five Bangor University students (27 females; M_{age} = 21.6, *SD* = 3.7), were recruited to participate via the University's participant panel, and received course credit as compensation. Of these participants, 27 (15 females; M_{age} = 22.4, *SD* = 4.8) received social feedback, and 28 (12 females; M_{age} = 20.9, *SD* = 1.9) received symbolic feedback during the learning phase of the experiment. Experimenters were blind to feedback condition when instructing participants in the learning phase. All participants provided written informed consent to take part in the study which was approved by the Ethics Committee of the School of Psychology at Bangor University. We excluded four individual datasets from all analyses because of excessive time spent on, or failure to complete, the learning phase (see criterion in Procedure below). We further excluded 4 datasets out of the remaining 51 from analysis for the learning phase and a different 4 out of 51 for the test phase; datasets were included on the basis of a sufficient number of trials to analyse being present. Thus, the final samples for statistical analysis in the learning phase was 47 (24 social, 13 females; 23 symbolic, 11 females) and 47 (23 social, 14 females; 24 symbolic, 11 females) for the test phase.

Stimuli

Eighty-one cards each featuring a unique combination of 1–3 shape(s) (circle, square, triangle), in one of three colours (red, green, blue), with one of three fillings (empty, hashed, full) were used to create card triads that either complied or not with the following rule: A legal combination is a triad of cards in which all cards are the same or different, considering each of the stimulus dimensions separately (shape, number, colour, filling). Any combination of cards featuring a partial repetition of any stimulus dimension was thus illegal (see Figure 1). Card triads were further split into 4 difficulty levels based on the perceived difficulty in assessing legality, e.g., a combination of cards failing the all same / all different criterion for all four dimensions was considered relatively easy to spot as illegal (cf. illegal difficulty level 1 in Figure 4.1).

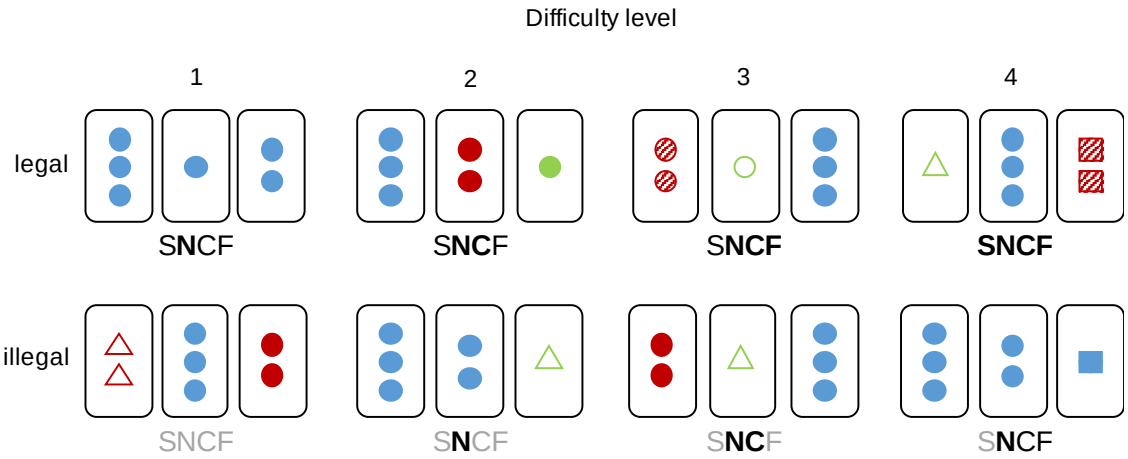


Figure 4.1. Examples of legal and illegal card triads split by levels of difficulty. Note that full repetition triads (same shape, S, number, N, colour, C, and filling, F) were not used in the experiment because they were too simple to identify. The code under each triad indicates the particular properties that comply with the rule: black slim letters code for a dimension repeated across all three cards, black bold letters code for dimensions different across all three cards, and grey letters indicate dimensions for which the all same / all different rule is violated.

The number of possible card combinations differed between difficulty levels (e.g., illegal level 1: 1 combination, illegal level 4: 32 possible combinations). In the learning phase, the weighting of each combination was adjusted to ensure that each level had equal probability of being presented throughout the staircase procedure in order to allow participants to learn about combinations from all of the levels. During the test phase,

however, we elected to present difficulty levels at their 'natural' frequency, and legal and illegal conditions were presented with a ratio of 1:3 to comply with the oddball design.

Card triads were presented under 4 degrees of visual angle in the learning phase and under less than 2 degrees of visual angle in the test phase, that is, in participants' foveal visual field so as to avoid eye saccades and consequent artefacts. Note that no ERPs were analysed in response to card triads in the learning phase.

Feedback: Twelve pictures of faces (6 female, 6 male) each presented with a happy, neutral or sad expression were collated from The Karolinska Directed Emotional Faces database (KDEF: Lundqvist, Flykt, & Öhman, 1998; Goeleven, De Raedt, Leyman, & Verschuere, 2008) and edited to fit within 2 degrees of visual angle. Six simple shapes (circle, square, triangle, hexagon, diamond, trapezoid) were drawn to fit the same surface as that covered by faces and coloured in green, orange or red in two different levels of luminance as a counterpart to the two genders for faces (6 dark, 6 light).

Procedure

Learning phase. Participants first completed an implicit-intentional learning task. On each trial, a card triad randomly selected from a database of card combinations was presented in the middle of a 19" CRT monitor with a refresh rate of 74 Hz. Legal and illegal combinations had equal probability of presentation as did levels of difficulty. Participants had to indicate whether the current combination was 'legal' or 'illegal' by pressing designated buttons on an SR response box (E-Prime 2.0 software; Psychology Software Tools, Pittsburgh, PA). Participants started on a random response basis. In the symbolic group, feedback was provided by means of shapes filled in one of two colours (green, correct; red, incorrect). Thus, the symbolic feedback stimuli shared some perceptual similarity with the card stimuli (i.e., some shapes and some colours) but the colour scheme of the symbolic feedback was semantically transparent (green for correct and red for incorrect) and binary,

thus entirely unambiguous whereas the shapes and colours presented on cards had no intrinsic meaning and were completely arbitrary. In the social group, participants received feedback from pictures of faces with one of two different emotional expressions (smile, correct; sad, incorrect).

Before the feedback stimulus, a neutral stimulus (orange shape in the symbolic group and neutral face in the social group) was displayed with a pseudorandom variable duration (750-950 ms in steps of 40 ms). The neutral stimulus served to focus the participant's attention in the centre of the screen (thus avoiding eye movements), and desynchronised the fERN response elicited by the subsequent valenced feedback stimulus from the ERP elicited by the card triad.

Participants progressed through a staircase procedure such that they had to make five correct cumulative judgements for each level of difficulty in each legal and illegal condition before triads from that level and condition were dismissed from training (see Figure 4.2). Any error reset the count of correct trials to zero for the current level of difficulty and condition. Response side was counterbalanced across participants.

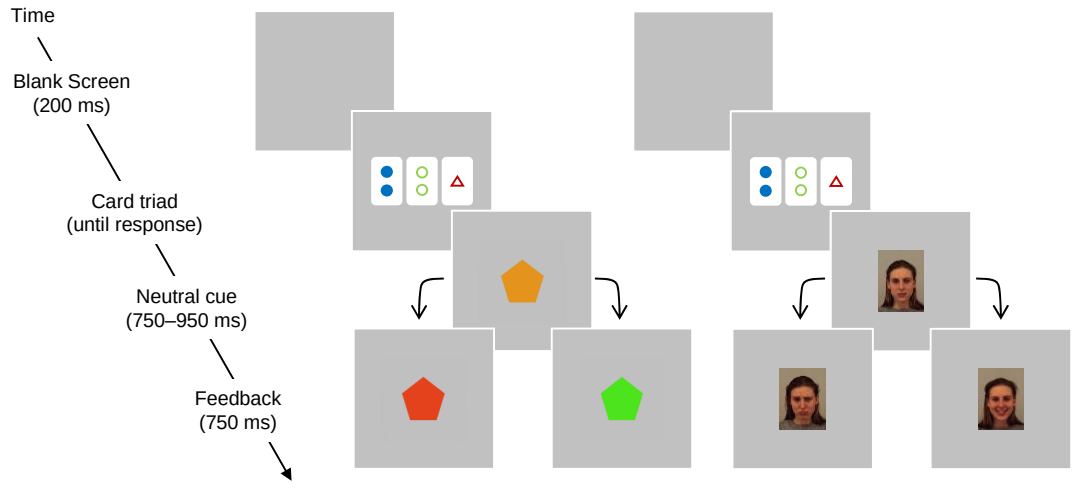


Figure 4.2. Structure of a trial on the staircase learning procedure**.** After presentation of the card triad, participants received one of two types of feedback: symbolic (left) or social (right), preceded by a neutral stimulus in all cases.

Testing phase. After completing the learning phase, participants were asked in a second phase to indicate whether each card triad presented was legal or illegal without feedback. Each triad was presented for a maximum duration of 1500 ms and response initiated the next triad presentation after an inter-stimulus interval of 450-550 ms (in steps of 25 ms) during which a fixation cross then appeared in the centre of the screen. The random inter-stimulus interval was introduced so as to reduce cross-trial ERP contamination (See Figure 4.3). There were three blocks of 200 trials and participants were given the chance to rest between each block. Legal trials were presented with an average frequency of 25% (range 23–26%) and thus were expected to act as deviants amongst frequent illegal triads, thus conforming to the structure of a classic oddball paradigm prone to eliciting P3b ERP effects.

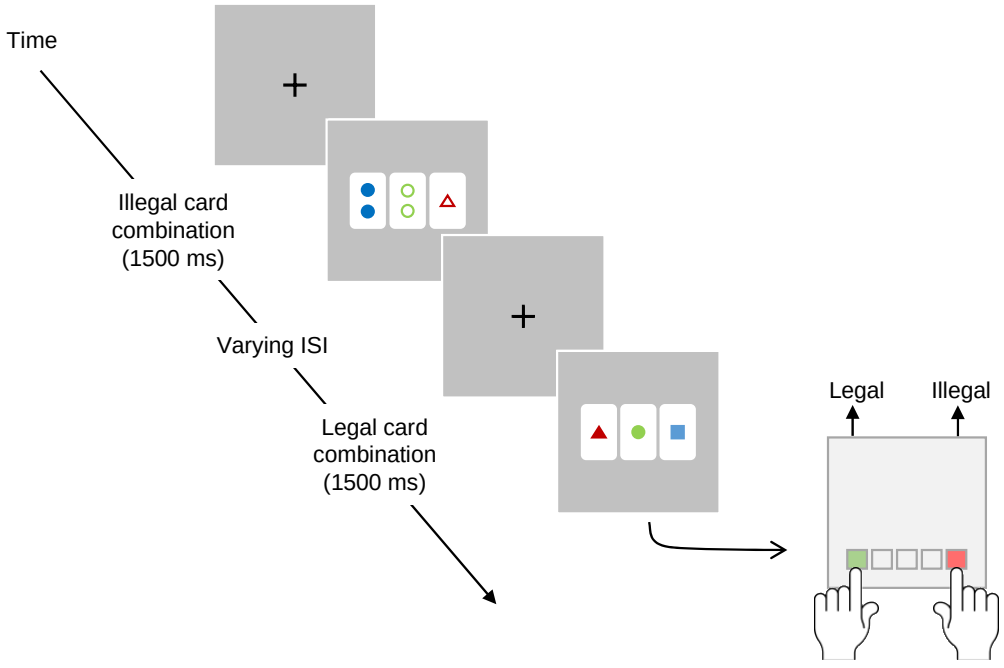


Figure 4.3. Trial structure of the test phase. Participants were required to respond on every trial using the SR box provided. Response side counterbalanced across participants.

EEG recording and analysis

EEG data were recorded continuously at a sampling rate of 1 kHz in reference to Cz using 64 Ag/AgCl electrodes attached to an elastic cap (Easycap^{TM}, Herrsching, Germany) and arranged according to the extended 10-20 convention. EEG signals were amplified using

SynAmps2™ (Neuroscan™ Inc., El Paso, Texas, USA). The ground electrode was placed at FPz. Four additional electrodes were placed to the right of the right eye and to the left of the left eye (HEOG) and above and below the right eye (VEOG) so as to monitor horizontal and vertical eye movements. Impedances were kept below 5 kΩ. Recordings were filtered on-line between 0.01 and 200 Hz (slope 24 db/Oct.).

The EEG data were filtered offline using a zero phase shift bandpass digital filter between 0.1Hz [24 db/Oct]–25 Hz [48 db/Oct] using Scan 4.5 (Neuroscan™ Inc., El Paso, Texas, USA). Major artefacts were manually rejected and eye blinks were mathematically corrected according to the procedure described in Gratton, Coles, and Donchin (1983). Continuous EEG activity was then segmented into epochs ranging from -100 to 1000 ms after stimulus onset for the learning phase and -200 to 1000 ms for the testing phase. A shorter baseline window was selected in the learning phase to minimise baseline contamination by the preceding neutral stimulus cue. Baseline correction was performed in reference to pre-stimulus activity, and individual averages were digitally re-referenced to the global average reference.

In the learning phase, the average number of feedback trials included in the symbolic condition was 43.48 (SEM = 4.64) and 45 (SEM = 8.20) in the social condition. As for the test phase, only accurate trials were kept for the analysis, leading to an average of 381.54 (SEM = 12.97) trials in the symbolic condition and 346.48 (SEM = 14.89) trials in the social condition.

The ERP modulation of interest in the learning phase was the feedback error-related negativity (fERN), which is typically maximal over frontocentral electrodes and typically peaks between 200-320 ms (Ma, Meng, & Shen, 2015; Miltner, Braun, & Coles, 1997; Scheffers & Coles, 2000). We thus analysed fERN mean amplitude at FC1, FCz, FC2, C1, Cz, and C2 between 200–320 ms, the predicted time-windows based on previous studies.

fERN peak latency was fixed in each condition and each participant, and measured at the electrode of minimum amplitude FCz where the fERN was most negative (See Picton et al., 2000). As for the test phase, rare legal stimuli were expected to elicit larger P3b amplitudes than frequent illegal stimuli. P3b mean amplitudes were analysed over the predicted centroparietal region (C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2) between 480-580 ms where it is classically analysed in tasks requiring elaborate cognitive processing (Kok, 2001; Polich, 2007).

Statistical analyses

Behavioural and ERP results were analysed using mixed design ANOVAs with legality (illegal, legal) as repeated-measures factors and feedback condition (social, symbolic) as a between groups factor. Greenhouse-Geisser corrections were applied when necessary, and *df* and *p* values reported are adjusted.

Results

Learning phase

Performance. Because speedy responses were not encouraged, we did not analyse reaction times in the learning phase. Analysis of accuracy showed a main effect of legality, $F(1,49)=15.93$, $p<.001$, $\eta_p^2=.25$, but no main effect of feedback condition or interaction (*p*-values ≥ .819). Participants responded to legal combinations significantly more accurately than illegal ones (M_{Legal} =.73, *SD* =.12; $M_{Illegal}$ =.66, *SD* =.10).

fERN. In order to analyse the fERN, a difference waveform was computed by subtracting the grand average positive feedback waveform from the negative feedback waveform (Miltner et al., 1997; Scheffers & Coles, 2000). There was no main effect of legality or feedback condition on mean fERN amplitude, nor significant interaction between the two (*p*-values ≥.228).

However, there was a significant main effect of feedback condition on fERN peak latency, F(1,45)=5.879, p=.019, η_p^2=.12. There was no main effect of legality or interaction between legality and feedback condition (p-values ≥.560). Therefore, we collapsed mean latencies across legality and tested feedback condition using an independent samples t-test and found a significant difference, t(45)=-3.08, p=.004, d=0.90 confirming the previous result (see Figure 4.4).

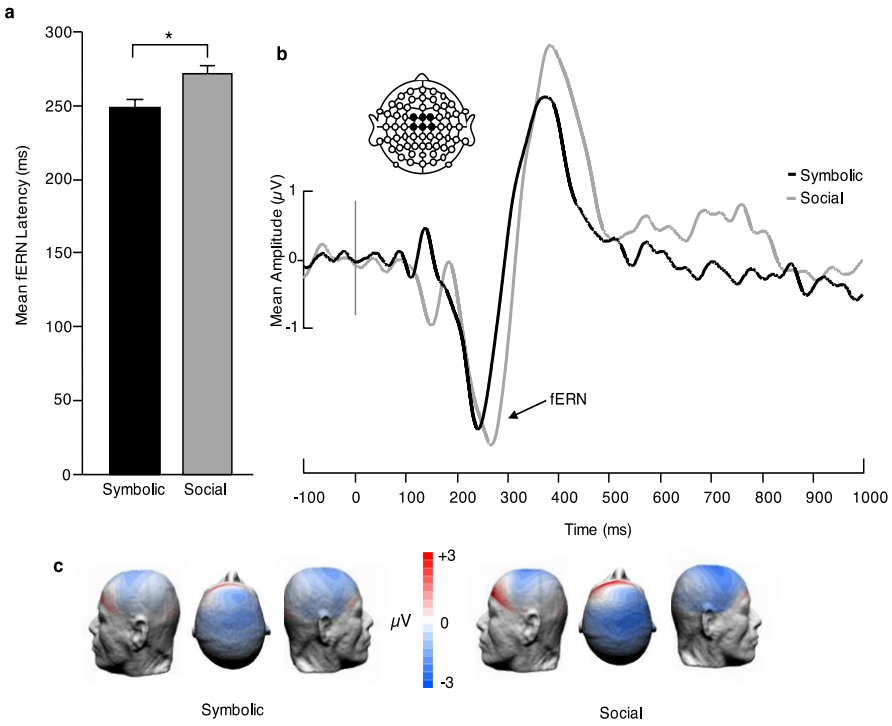


Figure 4.4. Effect of feedback condition on the fERN (negative minus positive feedback). (a) fERN mean latencies by feedback condition. (b) Grand-average ERP difference waveforms elicited over the frontocentral region (linear derivation of FC1, FCz, FC2, C1, Cz, and C2) in the symbolic (black line) and social (grey line) conditions. (c) fERN topographies (200-320 ms) by feedback condition. *p < .05.

Test phase

Performance. In terms of reaction times, there was a main effect of legality F(1,49)=6.89, p=.012, η_p^2=.12, with participants responding more quickly to illegal ($M_{Illegal}$ =786.92, SD =136.23) than legal cards (M_{legal}= 822.10, SD = 150.60). However, there was not a main effect of feedback condition, nor an interaction (p-values ≥ .538). Similarly with accuracy, we found a significant main effect of legality F(1,49)=4.33, p=.043, η_p^2=.08, but all

other effects were non-significant (*p*-values ≥ .588), with participants responding more accurately to illegal ($M_{Illegal}$ = .64, *SD* = .16) than legal cards (M_{legal} = .55, *SD* = .20), regardless of feedback condition (see Figure 4.5).

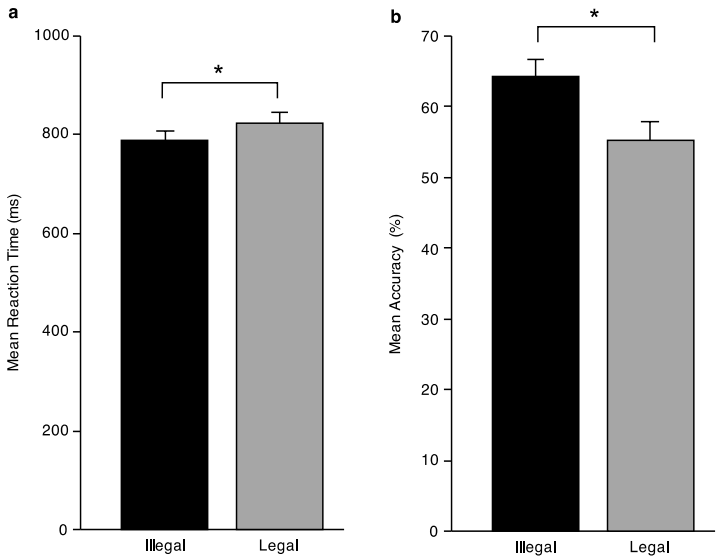


Figure 4.5. Behavioural results in the test phase - (a) reaction time and (b) accuracy, both collapsed across feedback condition. Error bars represent SEM. *p < .05.

P3b. There were significant main effects of legality, F(1,45)=30.29, *p*<.001, η_p^2=.40, and feedback condition, F(1,45)=10.49, *p*=.002, η_p^2=.19, on P3b mean amplitudes. There was also a significant interaction between the two, F(1,45)=11.04, *p*=.002, η_p^2=.20. A simple effects analysis showed that the difference in amplitude between illegal (standard) and legal (deviant) trials was significant in the symbolic, F(1,45)=39.80, *p*<.001, η^2=.46, but not the social feedback group, F(1,45)=2.33, *p*=.134, η^2=.03 (See Figure 4.6).

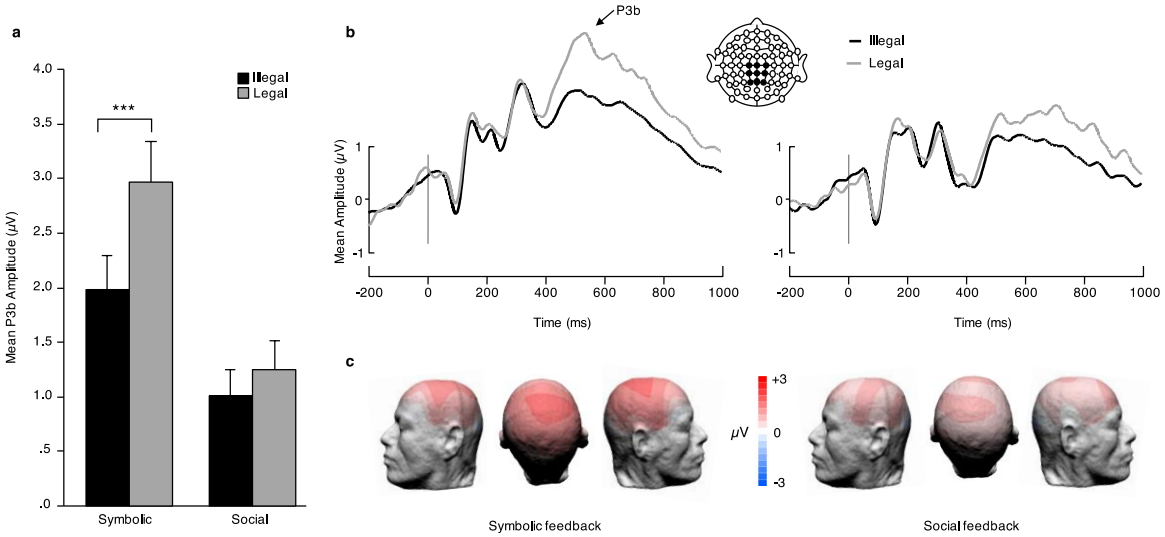


Figure 4.6. P3b Results. (a) Mean P3b amplitudes between 480-580 ms elicited over the centroparietal region (C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2) by legal (grey) and illegal (black) card triads in the symbolic (left) and social (right) feedback groups. Error bars indicate SEM. (b) P3b ERP waveforms elicited over the centroparietal region by legal (grey) and illegal (black) in the symbolic (left) and social (right) feedback groups. (c) P3b effect (legal minus illegal) topography by feedback group. ***p < .001.

Across testing phases: P3b and fERN

Finally, in a subset of participants (N = 43) whose datasets were retained for both learning and test phases, we predicted that as the latency of the fERN increases, the difference between the oddball and standard stimuli (P3b effect) would decrease. We suggest that a delay in processing hindered feedback registration in the learning phase, and lead to a weaker ERP response discriminating between illegal and legal stimuli at test. Analysis showed that there was indeed a significant negative correlation between these variables, $r(41)=-.27$, $p=.039$, $R^2=.07$ (1-tailed), in the absence of an interaction with group, $p > .1$ (See Figure 4.7).

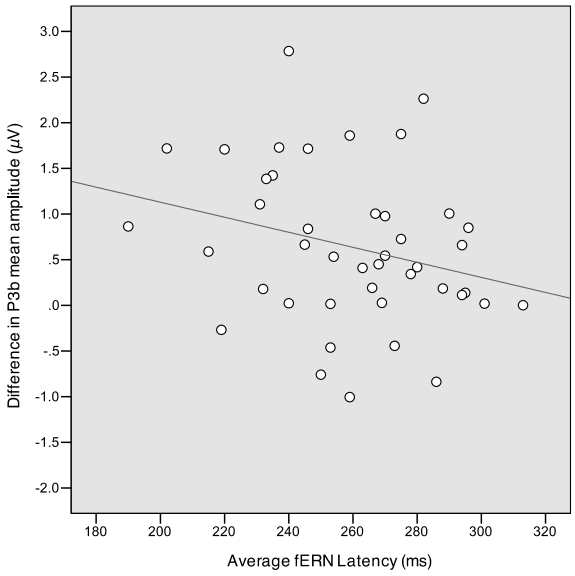


Figure 4.7. Relationship between fERN mean latency and P3b effect mean amplitude. Negative correlation between fERN mean latency and P3b effect (legal minus illegal) mean amplitude.

Discussion

In the present study, we compared two groups of participants learning a new card game and receiving two different types of feedback: symbolic or social. We investigated the effect of feedback type on performance during learning and at test, using behavioural and ERP measures. In the learning phase, participants were more accurate overall for legal than illegal card triads, regardless of feedback type. Although we did not find any difference in mean ERP amplitude between groups, the fERN elicited by socially salient stimuli was significantly delayed (by ~20 ms) in comparison to that elicited by symbolic feedback, regardless of card triad legality. At test, participants showed greater accuracy for illegal than legal card combinations, suggesting that learning in the training phase had taken place. Furthermore, when collapsing across legality, participants in both groups displayed similar levels of accuracy, meaning that feedback type did not cause measurable differences in general performance. This being said, ERP differences between groups did manifest at test. Even though the paradigm we used was a non-traditional P3b design (i.e., participants were required to respond to all stimuli) a typical P3b effect was elicited in response to infrequent legal compared to frequent illegal stimuli in the symbolic group, while the social group failed to show a mean amplitude difference between conditions in the same time window. Finally, we found a correlation between fERN peak latency in the learning phase and P3b mean amplitude at test.

We expected socially salient feedback to enhance task performance (as suggested by results obtained by Hurlemann et al., 2010), but failed to find behavioural differences between participant groups. Nonetheless, brain responses differed between groups in both the learning and the testing phase, with the symbolic group displaying earlier fERN peaking time than the social group, as well as a significant P3b effect. Thus, whereas both groups learned the rule of the game to a similar extent, ERP measures indicate that the quality of learning differed depending on feedback type. This interpretation is consistent with the finding of a

negative correlation between fERN latency and P3b amplitude: the greater the delay of negative feedback registration, the weaker the subsequent distinction between legal and illegal stimuli.

Thus, although feedback type does not differentially affect behavioural performance, socially relevant feedback appears to add cognitive noise affecting learning at the neurophysiological level. In that sense, the results echo those reported by Hu et al., (2015) who asked participants to sort 3-digit numbers according to arbitrary categories and gave feedback via either socially relevant (emotionally expressive faces) or symbolic (traffic light icons) stimuli. The authors found that social feedback impaired performance relative to symbolic feedback, and that those learning via social feedback performed at comparable levels of accuracy as those learning based on symbolic feedback only after receiving a dose of intranasal oxytocin. The authors reasoned that the relative disadvantage afforded by emotional faces could be due to this stimulus type not being a customary form of feedback in Chinese culture. In other words, Chinese participants find emotive human faces disruptive during learning. Since we found a similar effect in our ERP data, the relative advantage of symbolic feedback may extend to western cultures. This effect may be explained by facial stimuli increasing cognitive load during difficult tasks (as in Doherty-Sneddon et al., 2002), which is consistent with other studies showing improvements in accuracy when participants avert the gaze from socially relevant stimuli during cognitive tasks (Glenberg et al., 1998; Phelps, Doherty-Sneddon, & Warnock, 2006). This result also concurs with the classic finding that the presence of others during cognitively demanding tasks can be detrimental to task performance (Bond & Titus, 1983). It is noteworthy that we used 6 different identities in the social feedback condition, thus incurring stimulus variability to a greater extent than that involved in the symbolic feedback group, given that symbolic feedback only varied in basic geometrical properties and lightness. The relatively greater diversity of stimuli in the social

feedback version of the experiment may therefore have contributed to increasing the cognitive load in that condition and thus partly account for the pattern of difference found in the ERP data. Indeed, Hu et al. (2015) reported that using an emoticon instead of photographs of faces as feedback in the social feedback condition improves learning.

Note that the sharing of attributes between the shapes presented on cards and those used to provide feedback in the symbolic participant group could hardly account for this result. Symbolic feedback sometimes featured circles or squares, which could be green or red, attributes that could also be represented in some cards. Whereas the semantic value of the symbolic shapes was entirely based on colour, unambiguous (i.e., green = correct, red = incorrect), and binary in nature, the shape and colour attributes of shapes on cards were entirely arbitrary and only had value when considered across cards. And indeed, there was no detrimental effect of the overlap in attributes between card shapes and symbolic feedback shapes, thus not causing any measurable consequences in this study.

Our task deliberately engaged spatial and abstract-reasoning capabilities. Thus, it was cognitively demanding, which may explain why we did not find social feedback to have a facilitatory effect on task performance, but rather tend to cause shallower learning, indexed by P3b amplitudes. We provided electrophysiological evidence that socially salient feedback –when the task at hand is abstract and relatively complex– is less conducive to facilitating implicit learning, as evidenced by a delayed fERN. Participants in this feedback group also showed a reduced P3b effect, reflecting a decreased ability to distinguish between rare legal stimuli and more frequent illegal stimuli. Our results thus support recommendations that during difficult cognitive tasks, people should avoid looking at other individuals in order to increase task accuracy (Phelps et al., 2006). However, future research is needed to generalise this finding to other learning contexts (e.g., socially relevant tasks) and investigate whether it

is the social quality of the stimuli or its informativeness in the task context that affects learning quality on the neurophysiological level.

Indeed, in the present study we employed a type of negative social feedback different from that used in previous studies (Hu et al., 2015; Hurlemann et al., 2010; Mihov et al., 2010). The latter experiments used 'angry' faces as negative feedback in social conditions, while we elected to use 'sad' faces. One could argue that sadness does not convey negative feedback in response to an error in learning contexts as efficiently as frowning or anger. This may have resulted in a slightly different emotional context, thus limiting the validity of direct comparisons between studies.

Another point to note was the use of unequal number of possible combinations of the four shape dimensions per level of difficulty at test (see "Stimuli" section). In the learning phase, we ensured that the different combinations had equal probability of presentation to offer participants a chance to learn all possible combination types. However, during the test phase, we allowed the combinations to occur at their natural frequency. For example, there are many possible combinations of illegal trials in which the rule is violated for one stimulus dimension only as compared to the case of violations affecting all dimensions simultaneously. Future studies will determine whether the local frequency of each difficulty level has a measurable impact on detecting legal combinations amongst illegal ones. Indeed, once a rule is learned, it is unclear how the diversity of test stimuli affects performance since the criterion acquired during the learning phase is binary in nature.

Conclusion

This study sought to investigate the role of feedback during an implicit-intentional learning task. Contrary to our hypotheses drawn from previous behavioural studies, we found that symbolic feedback was more effective than social feedback, as demonstrated by a delayed fERN in the social group during learning, and lower ability to distinguish between

card combinations that complied to an implicitly learnt rule and those that did not at test, as indexed by a neurophysiological index of target detection (P3b). We suggest that the social salience of feedback may interfere with the learning process, at least when the rule to be learnt is abstract and relatively complex. Such effects need to be further investigated in experiments directly manipulating task complexity in social vs. non-social feedback contexts, and characterise the importance of the type of social feedback received.

Chapter 5

General Discussion

Summary

This thesis sought to examine how exposure to social information affects decision-making. Specifically, we examined how socially-relevant punishments and rewards compare to non-social alternatives, in terms of their effects on cooperative behaviour and the ability to implicitly learn new rules. In the first two empirical chapters, we were particularly interested in how social and non-social punishments played a role in decisions to cooperate in a naturalistic social interaction context relative to a typical anonymised laboratory one; as well as whether social rewards and punishments were more effective than monetary equivalents. In the third empirical chapter, we wanted to investigate whether participants could implicitly learn a rule more efficiently when they received social versus non-social feedback and whether the neural signature of learning would vary between feedback conditions.

Broadly, this thesis found that the inclusion of socially relevant information has mixed effects on decision-making. During our unique version of the Public Goods Games, we found that incentives endangering individuals' social concerns (i.e., social punishments) increased cooperation in socially-enriched environments, relative to monetary sanctions. However, social rewards did not appear to have the same effect, suggesting that rewards and punishments in the Public Goods Game had differential effects on contribution behaviour depending upon the environment in which they are administered. Additionally, in these face-to-face environments, positive social interactions, characterised by higher amounts of shared positive affect, generally helped to drive contributions. Thus, the first two empirical chapters suggest that information from the social environment, as well as manipulating social incentives, has an effect on cooperative behaviour. This has implications for the way in

which the Public Goods Game is conceptualised within laboratory contexts (see discussion below).

At the intrapersonal level, we compared the efficacy of social feedback (that included social rewards in the form of smiles) and non-social symbolic feedback during an implicit learning task. We found that social feedback altered the neural signature of learning, although this did not manifest in differences in behavioural performance across feedback groups. We tentatively suggest that social feedback interfered with learning on a neurophysiological level (see below).

Main findings & contribution

Chapter 2. In the first empirical chapter, we were interested in looking at cooperative behaviour in the public goods game when punishment options were available. Typical laboratory versions of this game take place over computer terminals, meaning that interactions between other group members are normally anonymous (Fehr & Gächter, 2000a; Rand, Dreber, Ellingsen, Fudenberg, et al., 2009). In these contexts, monetary sanctions appear to be superior to punishments designed to emulate social disapproval, by assigning disapproval 'points' to defecting players (Masclet et al., 2003; Noussair & Tucker, 2005). However, these results appear contradictory to field studies that suggest that interventions manipulating an individual's social concerns are more consistently effective at maintaining cooperation, compared to cost/benefit manipulations (Kraft-Todd et al., 2015).

Indeed, in the real social world, people often directly communicate disapproval to partners who have behaved poorly (Gächter & Fehr, 1999), and the opportunity for these kinds of face-to-face interactions appears to relate to increases in cooperative behaviour (Balliet, 2009; Drolet & Morris, 2000; Ostrom, 2000). Thus, in socially impoverished laboratory contexts, where interaction with social partners is not possible, the "sting" of

social disapproval may be reduced, as there are no tangible social consequences for punished players (Masclet et al., 2003; Noussair & Tucker, 2005).

Thus, in this Chapter, we used a unique variant of the Public Goods Game (PGG) to address the issues outlined above. Here, we allowed participants to engage in naturalistic social interactions with other group members (face-to-face context), contrasted against a typical 'anonymous' game setting. We were interested in examining whether the sting of punishments (also referred to here as the 'cost' of punishment) changed depending on the game context and the punishment type available. We predicted that in the typical anonymised version of the game, Monetary (reducing the payoff of the punished player by 50%), but not Social punishment (publicly naming the lowest contributor) would serve to maintain contribution levels, thereby replicating previous findings.

Results from this chapter suggested that in anonymised interactions, monetary sanctions distributed to the lowest contributors each round (i.e. free-riders) were more effective than social punishments. However, during the face-to-face game, this pattern reversed, and social punishments (i.e. publicly naming the lowest contributor) became the most effective sanction type. Thus, we appeared to replicate common findings in the anonymous version of the game that suggest that monetary sanctions are the most effective sanction type (Fehr & Gächter, 2000a). However, in the face-to-face version of the game, we found that socially relevant punishments were the most effective sanction type, thereby improving upon previous 'social' punishments (i.e. 'disapproval points'; Masclet et al., 2003; Noussair & Tucker, 2005). This finding broadly concurs with field research showing that engaging an individual's social concerns is more consistently effective than cost-benefit manipulations (Kraft-Todd et al., 2015). One such example of a social concern is one's reputation and identity, since when these are at stake, behaviour tends to become more

cooperative (Andreoni & Petrie, 2004; Ariely et al., 2009; Engelmann & Fischbacher, 2009; Gächter & Fehr, 1999; Nowak & Sigmund, 1998a).

Additionally, those receiving a social punishment in the face-to-face setting raised their contributions significantly compared to their standard rate. This suggests that players were prepared to incur a "cost" to themselves to avoid receiving a social punishment again in future rounds. However, the receipt of a monetary punishment did not raise contributions significantly, suggesting that such punishments do not have the same 'sting' as social punishments in face-to-face interactions. However, in the anonymous condition, players raised their contributions by similar amounts, regardless of punishment type.

Interestingly, in the face-to-face condition, we found that although players in all punishment conditions experienced similar amounts of interaction positivity (amount of shared positive affect, as rated by independent raters) on average. Furthermore, round-by-round differences in interaction positivity predicted players' contributions on the next round in conditions involving social punishments. Thus, direct social feedback from group members (i.e., laughter, friendly teasing) may account for the effectiveness of social punishments in this interaction setting.

Our research lends support to the idea that socially relevant information (via interaction with social partners) affects cooperative decision-making (Centorrino et al., 2015) and we further demonstrate that the interaction environment alter the effectiveness of certain punishment types. In anonymous interaction conditions, social sanctions do not appear to maintain cooperation over time. However, in enriched social contexts, social punishments are especially costly, even under conditions in which no information regarding other players' contributions is available. The immediacy of the social environment and threat to reputation present in this condition probably accounts for these results.

The efficacy of punishments is often studied alongside that of reward (for a meta-analytic review, see Balliet, Mulder, & Van Lange, 2011). Although research consistently suggests that punishment incentives are effective at maintaining cooperation in the public goods game, compared to when this mechanism is not in place (Fehr & Gächter, 2000a; Nowak & Sigmund, 1998b; Ostrom et al., 1992; Rand, Dreber, Ellingsen, Fundenberg, et al., 2009), it has also been suggested that costly punishment can be detrimental to group payoff (Dreber et al., 2008; Egas & Riedl, 2008; Szolnoki & Perc, 2010). However, it appears that the opportunity to reward players may mean that group earnings are healthier (Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). We thus explored the efficacy of rewards in the PGG in the next empirical chapter.

Chapter 3. In the second empirical chapter, we used the same public goods game paradigm, and looked at the effect of rewards (instead of punishments) on cooperative behaviour. Research has suggested that players are prepared to reward other generous players, even at a cost to themselves (Almenberg et al., 2011). However, an important requirement of this is that interactions are repeated (Walker & Halloran, 2004), as this allows players to develop a reputation over game rounds, which helps to maintain cooperation (Fu et al., 2008; Sigmund et al., 2001). The acquisition of a good reputation may even be a form of reward in itself, as known cooperators are often the beneficiaries of cooperation from others both within and outside of their own social groups (Semmann et al., 2005). In this experiment, we extended this idea, and compared the efficacy of rewards that benefitted the reputation of cooperators, to monetary incentives. We also compared these incentive types across two different interaction settings. We expected that in anonymous environments social/reputational rewards would be less effective than monetary, and in the face-to-face version of the game, we predicted that the enhanced reputational benefits of social rewards would increase cooperation relative to monetary rewards.

This set of experiments found that in anonymous settings, reward type (monetary, social or combined) did not differentially affect contributions behaviour. Thus, it did not appear that any one reward type was more effective than others at maintaining contributions to the public good in this interaction setting. However, in the face-to-face setting, we found the opposite effect to our prediction; monetary rewards were much more influential in terms of their contributions to investment behaviour than were social rewards.

In general, during both versions of the game, players were willing to pay a "cost" to distribute a reward to the highest contributor. On the surface perhaps, this seems to concur with other research showing that players are indeed willing to altruistically reward cooperative players (Almenberg et al., 2011). However, due to the way that we implemented the reward mechanism, players who decided to distribute a reward could receive their own reward back; thus, suggesting that players actually engaged in reward seeking behaviour. However, the exception to this was the Social reward condition in the face-to-face setting, in which participants did not pay more to distribute a reward. Additionally, in the face-to-face setting, positive social interactions predicted contributions in conditions where Monetary incentives were available, but not in the Social reward condition – here, players avoided raising contributions despite positive interactions.

Given that overall contributions were lower for the Social reward condition in the face-to-face setting and that this was the only condition in which participants were not willing to pay more to distribute a reward (perhaps to avoid receiving the reward themselves), we suggest that our British sample were reluctant to be singled out with these reputational rewards. These incentives were so explicit that they may have 'crowded-out' cooperation, in that they may have signalled selflessness for self-interested reasons (Ariely et al., 2009; Bénabou & Tirole, 2006).

For the first time, we have examined the effect of face-to-face social interaction in a cooperation game alongside reward incentives. Although the results were in some cases surprising (we did not anticipate that social rewards would be ineffective at raising contribution levels in face-to-face interactions), they did suggest that positive social interactions generally predict cooperative behaviour. This finding seems to concur with research showing that positive social signals from partners (i.e., smiles) tend to increase cooperative decisions (Scharlemann et al., 2001).

Thus, the first two chapters examined the effect of socially relevant rewards and punishments on an interpersonal level. These face-to-face environments allowed for the exchange of social cues between group members, which appeared to alter the efficacy of rewards and punishments, and thus to influence decision-making. In the next chapter, we examine the role of socially relevant feedback on an intrapersonal level, as this type of feedback is important for both predicting another's behaviour (Heerey & Velani, 2010) and adjusting one's own behaviour according to subtle shifts in the social environment (Kringelbach & Rolls, 2003). Social cues also provide information on how to maximise reward and minimise punishment (Heerey, 2014). Furthermore, socially relevant stimuli such as smiles, appears to be intrinsically more valuable than non-social feedback stimuli, even when both types of stimulus carry the same objective value (Heerey, 2014; Shore & Heerey, 2011). It also appears that social feedback is more beneficial for associative learning than non-social feedback (Hurlemann et al., 2010). However, it is unclear whether the reward value of socially relevant feedback confers benefits during learning, on an intrapersonal level, in tasks where instructions are not explicit. We explored this idea in the final empirical chapter.

Chapter 4. In this chapter, we examined whether socially-relevant feedback facilitates learning at an individual level by examining the neural signature of implicit

learning. It is well established that humans can implicitly learn contingencies about their environment without conscious awareness (Cleeremans et al., 1998; Dienes et al., 1995; Reber, 1967). However, the question of whether and how the social relevance of feedback affects the quality of learning is an important aspect that appears to have been neglected in the field of implicit learning.

We learn about our environment from other people, and this social information is abundant in everyday interactions. Indeed, research has shown that humans can implicitly learn cues from others to predict their future behaviour (Heerey & Velani, 2010). Similarly, in terms of the use of socially relevant feedback in non-social tasks, (e.g., in an associative task), social feedback has been shown to be more effective than non-social, symbolic feedback in the form of traffic light icons (Hurlemann et al., 2010). Given the high social relevance of human faces, we predicted that using this feedback stimulus would enhance performance during an implicit learning task compared to a non-social, symbolic stimulus (traffic lights). Participants were asked to learn which combination of visually presented cards conformed to a rule (legal) and which did not (illegal).

During the learning phase, participants were more accurate overall for legal than illegal card triads, regardless of feedback type. However, ERP results showed that the fERN was significantly delayed when participants received social rather than symbolic feedback. At test, we found that participants responded more quickly and accurately to illegal than legal card stimuli, but there were no differences between feedback conditions. However, it did appear that a P3b wave was elicited in response to rare legal stimuli in the symbolic group, a modulation not found in the social feedback group, suggesting that this group of participants were less able to discern between legal and illegal stimuli.

For the first time, we have examined the effectiveness of socially relevant and symbolic feedback during an implicit learning task as compared to non-social symbolic

feedback. Surprisingly, social feedback did not facilitate learning as indexed by performance, as we had predicted based on previous research showing that social feedback provides an advantage during an associative task (Hurlemann et al., 2010). This study suggests that social and symbolic feedback differentially affect neurophysiological activity, perhaps due to additional cognitive noise or increased competition for attention resources in the social feedback condition.

Implications

In the first two chapters, we examined how naturalistic social interactions affect cooperation during decision-making in the public goods game. Typical laboratory studies in this area use an anonymised methodology in which participants playing economic games cannot see or interact with other group members (Fehr & Gächter, 2000a; Masclet et al., 2003; Noussair & Tucker, 2005; Rand, Dreber, Ellingsen, Fundenberg, et al., 2009). Some research has started to consider the role of social information and cues from others during cooperative decision-making (e.g., Scharlemann et al., 2001; Schug et al., 2010). However, the social stimuli used are mostly static pictures of faces portraying a cue of interest (e.g., smiles; Scharlemann et al., 2001), or participant's facial expressions are video recorded for analysis in the absence of social partners (e.g., Schug, Matsumoto, Horita, Yamagishi, & Bonnet, 2010). Such methodological choices have been guided by the requirement of experimental control, but they considerably limit our ability to study real world cooperative decision-making mechanisms.

For this reason, we chose to allow participants to interact in a naturalistic way with other group members when they made decisions to cooperate, so as to get closer to the context in which individuals make real decisions. In the real world, we frequently make decisions within a social context, guided by the information gleaned by those around us. Indeed, we found that positive social interactions within a group generally help to drive

contributions towards the public good. Positive emotions, as identified in free responses made by participants playing an anonymous cooperation game, have been previously linked with cooperative behaviour (Rand et al., 2015). Beyond this, we have shown that it is possible to experimentally measure group positivity as it arises, allowing us to capture a more realistic account of cooperative decision-making.

Our cooperation experiments have also shown that the effectiveness of incentives depend on the setting in which they are administered. In previous anonymous laboratory contexts, punishments that attempt to operationalise social disapproval appear to be less effective than monetary sanctions (Masclet et al., 2003; Noussair & Tucker, 2005). Here, we found that publicly naming low contributors in a face-to-face situation increases the sting of receiving a punishment. Our version of social punishment was perhaps a closer approximation of the cognitive-emotional situation as it is experienced in the real world. Intuitively, it seems that real world punishment often takes the form we implemented in our study, whether it be individuals calling out free-riders during group work settings, or the media naming politicians and bankers for self-interested behaviour. Thus, this work adds credibility to the idea that engaging participants' social concerns increases cooperation (Kraft-Todd et al., 2015), and it provides hints for a simple and effective solution to tackle free-riding in social environments. However, such an intervention must also be advocated with caution, especially when it is employed in online social environments, such as Twitter. The effects of being publicly named for perceived bad behaviour in this environment can be disproportionately severe (for anecdotal accounts, see Ronson, 2015).

We also found that people were reluctant to pursue social rewards in socially enhanced environments, whereas this was not the case in anonymous environments. Perhaps the explicit incentive of gaining a good reputation may have 'crowded out' cooperation (Ariely et al., 2009; Bénabou & Tirole, 2006; Kraft-Todd et al., 2015). This finding implies

that in situations requiring the contributions of others, i.e., the case of charitable organisations, cooperation should not be incentivised with overt reputational rewards, e.g., the public naming of large donors. Public naming may dilute the intended signal, making it unclear whether the actor is really prosocial or whether she is acting in this way so as to gain a good reputation (Ariely et al., 2009). Thus, here we present evidence that in social, face-to-face environments, this crowding out effect seems to occur and that explicit rewards may better be avoided in such situations.

Thus, when decisions take place in anonymous environments, as is typical in this field, we are missing out on a great deal of information (e.g., the reciprocity of certain social cues, or aspects of the social environment, that have the potential to shape behavioural responses) which may help to unpack real world cooperative behaviour. Thus, this research has important implications for the field, in that we have taken a first step in identifying some of these untapped social interaction factors. This is also important from an epistemological point of view, since empirical social science research must be able to make valid inferences about the underpinnings of real-world behaviour.

ERP results from the final empirical chapter somewhat counter-intuitively suggested that social feedback can alter or interfere with learning at a neurophysiological level, as compared to non-social, symbolic feedback. Given the premise that social feedback is more valuable than non-social feedback (Heerey, 2014; Shore & Heerey, 2011) and that it is conducive to increased learning during associative tasks (Hurlemann et al., 2010), our result suggests that social feedback has a context-dependent effect. In difficult abstract reasoning tasks, where individuals are learning without explicit instruction, symbolic feedback might provide a clearer, simpler signal affording less neurophysiological interference. This has potential implications for the design of learning environments that typically employ socially

relevant feedback such as schools. However, there is a need to investigate in more detail what the neurophysiological differences between feedback types mean for future task performance.

Limitations

As with much work in the field of experimental economics, the games that are played often involve only small stakes. Indeed, in our public goods games, players only gambled with 10 pence at a time. Thus, we do not know whether the effectiveness of monetary punishments/rewards would change had the stakes been higher. It may be the case that larger monetary sanctions increase the efficacy of these punishments, regardless of interaction setting, but this cannot be assessed within the realms of our investigation.

It is important to also note that face-to-face interactions are difficult and time consuming to code in detail. Here, we instead asked independent observers to rate the general positively-valenced affectivity of the interactions, which was then condensed into a 'positivity' factor. There may be other, subtler social phenomena occurring during these interactions that our measures did not capture, which may be equally predictive of cooperative behaviour. Nonetheless, we must bear in mind that our experiments are a first step in exploring how information transmitted during social interactions may affect cooperation.

Finally, in terms of social feedback and implicit learning, there may have been some ambiguity in the social feedback that we used to indicate incorrect responses. We used sad faces as negative feedback and such facial expression does not unequivocally convey that an error has been committed. There was naturally less ambiguity with the traffic light feedback. In real life interactions, we are more likely to see frowning expressions when errors occur, and so, we do not know how ambiguity in face expression *per se* may have affected the results. Thus, in future experiments it may be productive to use a more realistic social cue that indicates an error has been made.

Future directions

Here, we have added a unique element to the public goods game paradigm by allowing concurrent face-to-face interactions with social partners. Future research in the field should build on this trend by continuing to explore what defection, cooperation, reward and punishment behaviours look like in more realistic interaction settings and how these may be realistically captured in the laboratory. Since the propensity for humans to engage with costly punishment behaviour may be an 'real artefact' of the laboratory (Guala, 2012), and even though our experiments are a step in the right direction, there is still some way to go in order to fully characterise the parameters underpinning real world behaviour. Also, we will need to investigate in more depth the role of social information from interaction partners during cooperative decision-making. I am particularly interested in what the social environment looks like in the lead up to a non/cooperative decision. How are these decisions mediated by the cues that we receive from others? Here, we have found that group interaction positivity is generally predictive of cooperation, and thus future studies should be able to identify the detailed characteristics of these interactions between group members; and the specific social cues that accompany them. Perhaps the incidence and reciprocity of smiles, examined using the Facial Expression Coding System (Kring & Sloan, 2007), would be a valid starting point for future work.

Furthermore, I am also interested in the interaction between cooperative decision-making and the perception of personal characteristics of interaction partners. How does the perception of others affect the propensity for cooperation, and the willingness to punish or reward others? To investigate this, I would start by analysing interpersonal ratings of other group members' personality characteristics using the Social Relations Model of social interaction (Back & Kenny, 2010) and associated analyses (i.e. the Round Robin analysis; Schönbrodt, Schmukle, & Back, 2011; Warner, Kenny, & Stoto, 1979). This will allow the description of dyadic processes between 'actors' and 'partners', as well as their unique

'relationship'. Such an approach should help resolve the issue of interdependence in the data when individuals interact within dyads or groups (Heerey, 2015) and may help shed light on how the perceived characteristics of group members affect cooperation, after interactions have occurred.

In terms of social feedback and implicit learning, we need to determine whether the neurophysiological differences observed between learning contexts using different types of feedback transpire into behavioural performance and information retention at a later stage. It may also be fruitful to examine whether the neural signature of feedback type, and perhaps even online behavioural performance, differs depending on the social relevance of the task, as our task was abstract and bore few real social ramifications. Additionally, considering previous research showing that reinforcement value differs between types of social rewards (e.g., Shore & Heerey, 2011), it may be interesting to examine whether different types of social feedback (e.g., genuine versus polite smiles) specifically affect implicit learning on behavioural and neurophysiological levels.

Conclusions

Overall, the work presented in this thesis shows that exposure to social information affects cooperative decision-making during the Public Goods Game, and alters the neural signature of implicit learning. In the Public Goods Game, the efficacy of rewards and punishments is contingent on both the form that these incentives take and the environment in which they are administered. Here, we have taken a first step towards more naturalistic testing, by examining how concurrent face-to-face social interactions between group members affect cooperative decision-making. We found that group interaction positivity generally helps drive contributions. These findings have notable consequences regarding the way in which this economic game is conceptualised within laboratory contexts, because many models of human cooperative behaviour fail to account for the fact that people interact

with others when making the decisions in the real world. This also has implications regarding the form that rewards and punishments should take in such experiments, depending on the context in which they are employed, to effectively promote cooperation. We also found that socially-relevant feedback, including social rewards in the form of smiles, somewhat counter-intuitively interfere with implicit learning at the neurophysiological level. Overall, this thesis serves to initiate a few new steps towards understanding how the social context in which people find themselves may influence both explicit and implicit aspects of decision-making. I believe that capturing real world mechanisms of social interactions during cooperative decision-making, not only in terms of externally observable manifestations but also in terms of unconscious, implicit mechanisms underlying them, is the future of the social and cognitive sciences.

References

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Almenberg, J., Dreber, A., Apicella, C. L., & Rand, D. G. (2011). Third party reward and punishment: Group size, efficiency and public goods. In N. M. Palmetti (Ed.), *Psychology of Punishment*. Nova Science Publishers.

Andreoni, J. (1988). Why free ride? *Journal of Public Economics*, *37*, 291–304. Retrieved from https://doi.org/10.1016/0047-2727(88)90043-6

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, *85*(4), 891–904. Retrieved from http://www.jstor.org/stable/2118238

Andreoni, J., & Croson, R. T. A. (2008). Partners versus Strangers: Random Rematching in Public Goods Experiments. In *Handbook or Experimental Economics Results, Volume 1* (Vol. 1, pp. 776–783). Elsevier B.V. https://doi.org/10.1016/S1574-0722(07)00082-0

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2003). the carrot or the stick: rewards, punishments and cooperation. *The American Economic Review*, *93*(3), 893–902. Retrieved from http://www.jstor.org/stable/3132122

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, *103*(418), 570–585. https://doi.org/10.2307/2234532

Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, *88*(7–8), 1605–1623. https://doi.org/10.1016/S0047-2727(03)00040-9

Arechar, A., Dreber, A., Fudenberg, D., & Rand, D. (2017). I'm just a soul whose intentions are good: The role of communication in noisy repeated games. *Games and Economic Behavior*, *104*, 726–743. https://doi.org/10.1016/j.geb.2017.06.013

Ariely, D., Bracha, A., & Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *The American Economic Review*, *99*(1), 544–555. Retrieved from http://www.jstor.org/stable/29730196

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Averbeck, B. B., & Duchaine, B. (2009). Integration of social and utilitarian factors in decision making. *Emotion (Washington, D.C.)*, *9*(5), 599–608. https://doi.org/10.1037/a0016509

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books (AZ).

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396.

Back, M. D., & Kenny, D. A. (2010). The social relations model: How to understand dyadic processes. *Social and Personality Psychology Compass*, *4*(10), 855–870. https://doi.org/10.1111/j.1751-9004.2010.00303.x

Baldwin, K. B., & Kutas, M. (1997). An ERP analysis of implicit structured sequence learning. *Psychophysiology*, *34*(1), 74–86. https://doi.org/10.1111/j.1469-8986.1997.tb02418.x

Balliet, D. (2009). Communication and Cooperation in Social Dilemmas: A Meta-Analytic Review. *Journal of Conflict Resolution*, *54*(1), 39–57. https://doi.org/10.1177/0022002709352443

Balliet, D., Mulder, L. B., & Van Lange, P. M. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the '"tragedy of the commons."' *Evolution and Human Behaviour*, *25*, 209–220. https://doi.org/10.1016/j.evolhumbehav.2004.04.002

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings. Biological Sciences / The Royal Society*, *274*(1610), 749–753. https://doi.org/10.1098/rspb.2006.0209

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad Is Stronger Than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Baumol, W. J. (2004). Welfare Economics and the Theory of the State. In *The Encyclopedia of Public Choice* (pp. 937–940). Boston, MA: Springer US. https://doi.org/10.1007/978-0-306-47828-4_214

Bayliss, A. P., & Tipper, S. P. (2005). Gaze and arrow cueing of attention reveals individual differences along the autism spectrum as a function of target context. *British Journal of Psychology*, *96*, 95–114. https://doi.org/10.1348/000712604X15626

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*, 245–249. https://doi.org/10.1038/nature07538

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior - Slides. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1017/CBO9781107415324.004

Bicchieri, C., & Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: an experimental analysis. *Politics, Philosphy & Economics*, *6*(2), 139–168. https://doi.org/10.1177/1470594X07077267

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, *60*(1), 11–26. https://doi.org/10.1016/j.jebo.2003.06.006

Bond, C. F., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, *94*(2), 265–292. https://doi.org/10.1037/0033-2909.94.2.265

Boyd, R., & Richerson, P. J. (1992). Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology*, *13*, 171–195. https://doi.org/https://doi.org/10.1016/0162-3095(92)90032-Y

Burnham, T. C., & Hare, B. (2007). Engineering Human Cooperation Does Involuntary Neural Activation Increase Public Goods Contributions? *Human Nature*, *18*, 88–108. https://doi.org/10.1007/s12110-007-9012-2

Burton-Chellew, M. N., & West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, *110*(1), 216–221. https://doi.org/10.1073/pnas.1210960110

Burton-Chellew, Maxwell N., & West, S. A. (2012). Correlates of Cooperation in a One-Shot High-Stakes Televised Prisoners' Dilemma. *PLoS ONE*, *7*(4), e33344. https://doi.org/10.1371/journal.pone.0033344

Burton-Chellew, Maxwell N, Nax, H. H., & West, S. A. (2015). Payoff-based learning explains the decline in cooperation in public goods games. *Proceedings of the Royal Society*, *282*. https://doi.org/10.1098/rspb.2014.2678

Camerer, C. (1998). Bounded rationality in individual decision making. *Experimental Economics*, *1*(2), 163–183. https://doi.org/10.1007/BF01669302

Camerer, C. F. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. Retrieved from http://ssrn.com/abstract=1977749

Camerer, C., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science (New York, N.Y.)*, *311*(5757), 47–52. https://doi.org/10.1126/science.1110600

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2015). Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution and Human Behavior*, *36*(1), 8–16. https://doi.org/10.1016/j.evolhumbehav.2014.08.001

Chaudhuri, A. (2010). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47–83. https://doi.org/10.1007/s10683-010-9257-1

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5). https://doi.org/https://doi.org/10.1016/j.tics.2011.03.003

Chun, M. M., & Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Cialdini, R. B. (2003). Crafting Normative Messages. *Current Directions in Psychological Science*, *12*, 105–109. https://doi.org/10.1111/1467-8721.01242

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Sciences*, *2*(10), 406–416. https://doi.org/https://doi.org/10.1016/S1364-6613(98)01232-7

Clutton-Brock, T. H., & Parker, G. A. (1995). punishment in animal societies. *Nature*, *373*, 209–215. Retrieved from https://www.researchgate.net/profile/Geoff_Parker2/publication/15383342_Punishment_in_Animal_Societies/links/09e4151445c696a16e000000.pdf

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans

reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276. https://doi.org/https://doi.org/10.1016/0010-0277(89)90023-1

Cosmides, L., & Tooby, J. (1995). Cognitive Adaptations for Social Exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford University Press.

Costa, D. L., & Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, *11*(3), 680–702. https://doi.org/10.1111/jeea.12011

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The Value of Vengeance and the Demand for Deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279–2286. https://doi.org/http://dx.doi.org/10.1037/xge0000018

Croson, R. T. A. (1996). Partners and strangers revisited. *Economics Letters*, *53*(1), 25–32. https://doi.org/10.1016/S0165-1765(97)82136-2

Dal Bó, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *The American Economic Review*, *95*(5), 1591–1604. Retrieved from http://www.jstor.org/stable/4132766

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, *31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258. https://doi.org/10.1126/science.1100735

Dewall, C. N., Baumeister, R. F., Gailliot, M. T., & Maner, J. K. (2008). Depletion Makes the Heart Grow Less Helpful: Helping as a Function of Self-Regulatory Energy and Genetic Relatedness. *Personality and Social Psychology Bulletin*, *34*(12), 1653–1662. https://doi.org/10.1177/0146167208323981

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1322–1338. https://doi.org/10.1037/0278-7393.21.5.1322

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology*, *38*(3), 438–445. https://doi.org/10.1037/0012-1649.38.3.438

Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition*, *33*(4), 727–733. https://doi.org/10.3758/BF03195338

dos Santos, M., Rankin, D. J., & Wedekind, C. (2013). Human Cooperation Based On Punishment Reputation. *Evolution*, *67*(8), 2446–2450. https://doi.org/10.1111/evo.12108

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*, 348–351. https://doi.org/10.1038/nature06723

Drolet, A. L., & Morris, M. W. (2000). Rapport in Conflict Resolution: Accounting for How

Face-to-Face Contact Fosters Mutual Cooperation in Mixed-Motive Conflicts. *Journal of Experimental Social Psychology*, *36*, 26–50. https://doi.org/https://doi.org/10.1006/jesp.1999.1395

Duffy, J., & Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, *66*, 785–812. https://doi.org/10.1016/j.geb.2008.07.003

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of The Royal Society, Biological Sciences*, *275*, 871–878. https://doi.org/10.1098/rspb.2007.1558

Eimer, M., Goschke, T., Schlaghecken, F., & Stürmer, B. (1996). Explicit and implicit learning of event sequences: evidence from event-related brain potentials. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(4), 970–987. https://doi.org/10.1037/0278-7393.22.4.970

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehaiovural Reviews*, *24*, 581–604. https://doi.org/https://doi.org/10.1016/S0149-7634(00)00025-7

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behaviour*, *67*, 399–407. https://doi.org/10.1016/j.geb.2008.12.006

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. https://doi.org/10.1016/j.evolhumbehav.2010.10.006

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of Trustworthiness From Intuitive Moral Judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/http://dx.doi.org/10.1037/xge0000165

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third party punishment and social norms. *Evolution and Human Behaviour*, *25*, 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994. Retrieved from http://www.jstor.org/stable/117319

Fehr, E., & Gächter, S. (2000b). Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives*, *14*(3), 159–181. Retrieved from http://www.jstor.org/stable/2646924

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140. https://doi.org/10.1038/415137a

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137–140. https://doi.org/10.1038/nature01474

Fehr, E., & Rockenbach, B. (2004). Human altruism: economic, neural, and evolutionary

perspectives. *Current Opinion in Neurobiology*, *14*(6), 784–790. https://doi.org/10.1016/j.conb.2004.10.007

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9

Fiser, J., & Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467. https://doi.org/10.1037/0278-7393.28.3.458

Fu, F., Hauert, C., Nowak, M., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review*, *78*(2), 1–8. https://doi.org/10.1103/PhysRevE.78.026117

Gächter, S., & Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior & Organization*, *39*(4), 341–369. https://doi.org/10.1016/S0167-2681(99)00045-1

Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior & Organization*, *55*(4), 505–531. https://doi.org/10.1016/J.JEBO.2003.11.006

Gaertig, C., Moser, A., Alguacil, S., & Ruz, M. (2012). Social information and economic decision-making in the ultimatum game. *Frontiers in Neuroscience*, *6*, 1–8. https://doi.org/10.3389/fnins.2012.00103

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology*, *206*, 169–179. https://doi.org/10.1006/jtbi.2000.2111

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*, 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658. https://doi.org/10.3758/BF03211385

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition & Emotion*, *22*(6), 1094–1118. https://doi.org/10.1080/02699930701626582

Goldstein, D. G., & Gigerenzer, G. (2002). Models of Ecological Rationality: The Recognition Heuristic. https://doi.org/10.1037/0033-295X.109.1.75

Gratton, G., Coles, M. G. ., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Gravelle, H., & Rees, R. (2004). *Microeconomics* (3rd ed.). Edinburgh: Pearson Education Limited. Retrieved from http://eprints.whiterose.ac.uk/72753/

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *The Behavioral and Brain Sciences*, *35*(1), 1–15. https://doi.org/10.1017/S0140525X11000069

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The Competitive Advantage of Sanctioning Institutions. *Science*, *312*, 108–111. https://doi.org/10.1126/science.1123633

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037//0033-295X

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2014). Cooperating with the future. *Nature*, *511*, 220–223. https://doi.org/10.1038/nature13530

Hayashi, F. (2000). *Econometrics*. New Jersey: Princeton University Press.

Heerey, E. A. (2014). Learning from social rewards predicts individual differences in self-reported social ability. *Journal of Experimental Psychology: General*, *143*(1), 332–339. https://doi.org/10.1037/a0031511

Heerey, E. A. (2015). Decoding the dyad: Challenges in the study of individual differences in social behavior. *Current Directions in Psychological Science*, *24*(4), 285–291. https://doi.org/10.1177/0963721415570731

Heerey, E. A., & Crossley, H. M. (2013). Predictive and reactive mechanisms in smile reciprocity. *Psychological Science*, *24*(8), 1446–1455. https://doi.org/10.1177/0956797612472203

Heerey, E. A., & Velani, H. (2010). Implicit learning of social predictions. *Journal of Experimental Social Psychology*, *46*(3), 577–581. https://doi.org/10.1016/j.jesp.2010.01.003

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Mcelreath, R., … Gintis, H. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. In *Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Review* (Vol. 91, pp. 73–78). Retrieved from http://www.jstor.org/stable/2677736

Henrich, J., Mcelreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Henrich, N. (2006). Costly Punishment Across Human Societies. *Science*, *312*, 1767–1770. https://doi.org/10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367. https://doi.org/10.1126/science.1153808

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). preferences, property rights and anonymity in bargaining games. *Games and Economic Behaviour*, *7*, 346–380. https://doi.org/https://doi.org/10.1006/game.1994.1056

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of*

Sciences of the United States of America. https://doi.org/10.1073/pnas.1417904112

Hollander, S. (2000). *John Stuart Mill on economic theory and method*. *Collected essays III*. London and New York: Routledge. https://doi.org/10.4324/9780203439036

Hopfensitz, A., & Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, *119*(540), 1534–1559. https://doi.org/10.1111/j.1468-0297.2009.02288.x

Hu, J., Qi, S., Becker, B., Luo, L., Gao, S., Gong, Q., … Kendrick, K. M. (2015). Oxytocin selectively facilitates learning with social feedback and increases activity and functional connectivity in emotional memory and reward processing regions. *Human Brain Mapping*, *36*, 2132–2146. https://doi.org/10.1002/hbm.22760

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., … Kendrick, K. M. (2010). Oxytocin Enhances Amygdala-Dependent, Socially Reinforced Learning and Emotional Empathy in Humans. *Journal of Neuroscience*, *30*(14), 4999–5007. https://doi.org/10.1523/JNEUROSCI.5538-09.2010

Isaac, M. R., McCue, K. F., & Plott, C. R. (1985). Public goods provision in an experimental environment. *Journal of Public Economics*, *26*, 51–74. https://doi.org/https://doi.org/10.1016/0047-2727(85)90038-6

Isaac, M. R., & Walker, J. M. (1988). Communication and free-riding behaviour: the voluntary contribution mechanism. *Economic Inquiry*, *26*(4), 585–608. https://doi.org/10.1111/j.1465-7295.1988.tb01519.x

Isaac, M. R., Walker, J. M., & Williams, A. W. (1994). Group size and the voluntary provision of public goods - experimental evidence utilizing large groups. *Journal of Public Economics*, *54*, 1–36. https://doi.org/https://doi.org/10.1016/0047-2727(94)90068-X

Isaac, M. R., Walker, J., & Thomas, S. (1984). Divergent evidence on free-riding: An experimental examination of possible explanations. *Public Choice*, *43*(2), 113–149. https://doi.org/10.1007/BF00140829

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284–294. https://doi.org/10.1016/j.neuron.2008.03.020

Janacsek, K., & Nemeth, D. (2012). Predicting the future: from implicit learning to consolidation. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *83*(2), 213–221. https://doi.org/10.1016/j.ijpsycho.2011.11.012

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*. https://doi.org/10.1038/nature16981

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113/-/DCSupplemental

Jordan, J. J., & Rand, D. G. (n.d.). Building Costly Signaling from the Ground Up: A Model of Third-Party Punishment as a Costly Signal of Exposure to Repeated Interactions. *Journal of Theoretical Biology, Forthcoming*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Kenny, D. A., & La Voie, L. (1984). The Social Relations Model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (18th ed., pp. 142–182).

Kim, O., & Walker, M. (1984). The free rider problem: Experimental evidence. *Public Choice*, *43*(1), 3–24. https://doi.org/10.1007/BF00137902

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*(4), 826–842. https://doi.org/10.1037/a0011381

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557–577. https://doi.org/10.1017/S0048577201990559

Komorita, S. S., & Barth, J. M. (1985). Components of reward in social dilemmas. *Journal of Personality and Social Psychology*, *48*(2), 364–373. https://doi.org/10.1037/0022-3514.48.2.364

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. G. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, *3*, 96–101. https://doi.org/10.1016/j.cobeha.2015.02.006

Kring, A. M., & Sloan, D. M. (2007). The Facial Expression Coding System (FACES): Development, Validation, and Utility. *Psychological Assessment*, *19*(2), 210–224. https://doi.org/10.1037/1040-3590.19.2.210

Kringelbach, M. L., & Rolls, E. T. (2003). Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage*, *20*(2), 1371–1383. https://doi.org/10.1016/S1053-8119(03)00393-8

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior and Organization*, *76*, 225–237. https://doi.org/10.1016/j.jebo.2010.08.007

Landa, J. T., & Wang, X. T. (2001). Bounded rationality of economic man: Decision making under ecological, social, and institutional constraints. *Journal of Bioeconomics*, *3*(2–3), 217–235. https://doi.org/10.1023/A:1020597813814

Levitt, S. D., & List, J. A. (2007). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics*, *40*(2), 347–370. Retrieved from http://home.uchicago.edu/jlist/papers/99-fulltext.pdf

List, J. A., & Levitt, S. D. (2005). What Do Laboratory Experiments Tell Us About the Real World? Retrieved from http://pricetheory.uchicago.edu/levitt/Papers/LevittList2005.pdf

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF. ISBN 91-630-7164-9.

Ma, Q., Meng, L., & Shen, Q. (2015). You have my word: reciprocity expectation modulates feedback-related negativity in the trust game. *PloS One*, *10*(2), 1–10. https://doi.org/10.1371/journal.pone.0119129

Marwell, G., & Ames, E. (1981). Economists free ride, does anyone else? *Journal of Public Economics*, *15*, 295–310.

Marwell, G., & Ames, R. E. (1980). Experiments on the Provision of Public Goods. II. Provision Points, Stakes, Experience, and the Free-Rider Problem. *American Journal of Sociology*, *85*(4), 926. https://doi.org/10.1086/227094

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review*, *93*(1), 366–380. Retrieved from http://www.jstor.org/stable/3132181

Mihov, Y., Mayer, S., Musshoff, F., Maier, W., Kendrick, K. M., & Hurlemann, R. (2010). Facilitation of learning by social-emotional feedback in humans is beta-noradrenergic-dependent. *Neuropsychologia*, *48*(10), 3168–3172. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2010.04.035

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002a). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society*, *269*, 881–883. https://doi.org/10.1098/rspb.2002.1964

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002b). Reputation helps solve the "tragedy of the commons." *Nature*, *415*, 424–426. https://doi.org/10.1038/415424a

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback i n a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. Retrieved from http://www.krigolsonteaching.com/uploads/4/3/8/4/43848243/miltner1997.pdf

Molenmaker, W. E., De Kwaadsteniet, E. W., & Van Dijk, E. (2014). On the willingness to costly reward cooperation and punish non-cooperation: The moderating role of type of social dilemma. *Organizational Behavior and Human Decision Processes*, *125*, 175–183. https://doi.org/10.1016/j.obhdp.2014.09.005

Myerson, R. (1991). *Game theory: analysis of conflict*. Harvard University Press.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*, 91–112. https://doi.org/10.1016/j.jpubeco.2007.04.008

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics Econ*, *11*, 358–369. https://doi.org/10.1007/s10683-007-9171-3

Northover, S. B., Pedersen, W. C., Cohen, A. B., & Andrews, P. W. (2015). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior*, 1–10. https://doi.org/10.1016/j.evolhumbehav.2016.07.001

Noussair, C., & Tucker, S. (2005). Combining Monetary and Social Sanctions To Promote Cooperation. *Economic Inquiry*, *43*(3), 649–660. https://doi.org/10.1093/ei/cbi045

Nowak, M. A., & Sigmund, K. (1998a). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Nowak, M. A., & Sigmund, K. (1998b). The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology*, *194*, 561–574. https://doi.org/https://doi.org/10.1006/jtbi.1998.0775

Ostrom, E. (2000). Collective Action and the Evolution of Social Noms. *Journal of Economic*

Perspectives, *14*(3), 137–158. https://doi.org/http://dx.doi.org/10.1080/19390459.2014.935173

Ostrom, E., Walker, J. M., & Gardner, R. (1992). covenants with and without a sword: self governance is possible. *The American Political Science Review*, *86*(2), 404–417. https://doi.org/https://doi.org/10.2307/1964229

Parks, C. D., Joireman, J., & Van Lange, P. A. M. (2013). Cooperation, Trust, and Antagonism: How Public Goods Are Promoted. *Psychological Science in the Public Interest*, *14*(3), 119–165. https://doi.org/10.1177/1529100612474436

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647. https://doi.org/http://dx.doi.org/10.1287/mnsc.2015.2168

Phelps, F. G., Doherty-Sneddon, G., & Warnock, H. (2006). Helping children think: Gaze aversion and teaching. *British Journal of Developmental Psychology*, *24*(3), 577–588. https://doi.org/10.1348/026151005X49872

Picton, T. W., van Roon, P., Armilio, M. L., Berg, P., Ille, N., & Scherg, M. (2000). The correction of ocular artifacts: a topographic perspective. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *111*(1), 53–65. https://doi.org/10.1016/S1388-2457(99)00227-8

Polich, J. (2007). Updating P300: An Integrative Theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019.Updating

Potters, J., & Stoop, J. (2016). Do cheaters in the lab also cheat in the field? https://doi.org/10.1016/j.euroecorev.2016.03.004

Poundstone, W. (1992). Prisoner's Dilemma: John von Neumann, Game Theory, and the puzzle of the bomb (pp. 43–44). New York: Anchor Books.

Putnam, L. L., & Jones, T. S. (1982). the role of communication in bargaining. *Human Communication Research*, *8*(3), 262–280. https://doi.org/10.1111/j.1468-2958.1982.tb00668.x

Raihani, N. J., & Bshary, R. (2012). A positive effect of flowers rather than eye images in a large-scale, cross-cultural dictator game. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1742), 3556–3564. https://doi.org/10.1098/rspb.2012.0758

Raihani, Nichola J, & Bshary, R. (2015). The reputation of punishers. *CellPress*, *30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics & self-interested deliberation. *Psychological Science*, *27*(9), 1192–1206. https://doi.org/10.1177/0956797616654455

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(48), 19193–19198.

https://doi.org/10.1073/pnas.1108243108

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation (supporting online material). *Science (New York, N.Y.)*, *325*, 1272–1275. https://doi.org/10.1126/science.1177418

Rand, D. G., Dreber, A., Ellingsen, T., Fundenberg, D., & Nowak, M. (2009). Positive interactions promote public cooperation. *Science*, *325*, 1272–1275. https://doi.org/10.1126/science.1177418

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427–430. https://doi.org/10.1038/nature11467

Rand, D. G., Kraft-Todd, G., & Gruber, J. (2015). The Collective Benefits of Feeling Good and Letting Go: Positive Emotion and (dis)Inhibition Interact to Predict Cooperative Behavior. *PloS One*, *10*(1), 1–12. https://doi.org/10.1371/journal.pone.0117426

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(June), 3677. https://doi.org/10.1038/ncomms4677

Rand, D. G., Yoeli, E., & Hoffman, M. (2014). Harnessing Reciprocity to Promote Cooperation and the Provisioning of Public Goods. *Policy Insights from Behavioral and Brain Science*, *1*(1), 263–269. https://doi.org/10.1177/2372732214548426

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behaviour*, *6*(6), 855–863. https://doi.org/http://dx.doi.org/10.1016/S0022-5371(67)80149-X

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/http://dx.doi.org/10.1037/0096-3445.118.3.219

Reber, P. J., & Kotovsky, K. (1997). Implicit learning in problem solving: The role of working memory capacity. *Journal of Experimental Psychology: General*, *126*(2), 178–203. https://doi.org/10.1037/0096-3445.126.2.178

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics*, *9*, 79–101. https://doi.org/10.1007/s10683-006-4309-2

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Source: Proceedings: Biological Sciences*, *265*(1394), 427–431. Retrieved from http://www.jstor.org/stable/50853

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(45), 18307–18312. https://doi.org/10.1073/pnas.1108996108

Ronson, J. (2015). *So you've been publically shamed*. London: Picador.

Roth, A. A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and

Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *The American Economic Review*, *81*(5), 1068–1095. https://doi.org/10.1126/science.151.3712.867-a

Ruz, M., Moser, A., Webster, K., McCandliss, B., & Quartz, S. (2011). Social Expectations Bias Decision-Making in Uncertain Inter-Personal Situations. *PLoS ONE*, *6*(2), e15762. https://doi.org/10.1371/journal.pone.0015762

Saffran, J. R. (2003). Statistical Language Learning. *Current Directions in Psychological Science*, *12*(4), 110–114. https://doi.org/10.1111/1467-8721.01243

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928. https://doi.org/10.1126/SCIENCE.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *9*(70), 27–52. https://doi.org/https://doi.org/10.1016/S0010-0277(98)00075-4

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92. https://doi.org/10.1177/1043463195007001004

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*(5), 617–640. https://doi.org/10.1016/S0167-4870(01)00059-9

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 141–151. https://doi.org/10.1037/0096-1523.26.1.141

Schönbrodt, F. D., Schmukle, S. C., & Back, M. D. (2011). Round robin analyses in R : How to use TripleR, (April), 1–29.

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior*, *31*(2), 97–94. https://doi.org/10.1016/j.evolhumbehav.2009.09.006

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, *6*(1), 38–47. https://doi.org/10.1177/1745691610393528

Sefton, M., Shupp, R., & Walker, J. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, *45*(4), 671–690. https://doi.org/10.1111/j.1465-7295.2007.00051.x

Seger, C. A. (1994). Implicit Learning. *Psychological Bulletin*, *115*(2), 163–196. https://doi.org/http://dx.doi.org/10.1037/0033-2909.115.2.163

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review*, *50*, 581–602. https://doi.org/10.1016/j.euroecorev.2004.10.005

Semmann, D., Krambeck, H.-J., & Milinski, M. (2005). Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology*, *57*, 611–616.

https://doi.org/10.1007/s00265-004-0885-3

Shanks, D. R. (2005). Implicit Learning. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of Cognition* (pp. 202–220). SAGE publications.

Shanks, D. R., & Channon, S. (2002). Effects of a secondary tasks on "implicit" sequence learning: learning or performance? *Psychological Research, 66,* 99–109. https://doi.org/10.1007/s00426-001-0081-2

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–447. https://doi.org/10.1017/S0140525X00035032

Shore, D. M., & Heerey, E. A. (2011). The value of genuine and polite smiles. *Emotion, 11*(1), 169–174. https://doi.org/10.1037/a0022601

Shore, D. M., & Heerey, E. A. (2013). Do social utility judgments influence attentional processing? https://doi.org/10.1016/j.cognition.2013.06.011

Sigmund, K., Hauert, C., Nowak, M. A., & Wachter, K. W. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences, 98*(19), 10757–10762. https://doi.org/doi/10.1073/pnas.161155698

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466–469. https://doi.org/10.1038/nature04271

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences, 104*(44), 17435–17440. Retrieved from http://www.pnas.org/content/104/44/17435.full.pdf

Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies, 77,* 1540–1566. https://doi.org/10.1111/j.1467-937X.2010.00608.x

Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters, 6,* 659–662. https://doi.org/10.1098/rsbl.2010.0209

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,* 201–206. https://doi.org/10.1016/j.evolhumbehav.2012.11.009

Szolnoki, A., & Perc, M. (2010). Reward and cooperation in the spatial public goods game. *EPL (Europhysics Letters), 92,* 38003p1-38003p6. https://doi.org/10.1209/0295-5075/92/38003

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences, 104*(30), 1250–12535. https://doi.org/doi/10.1073/pnas.0609927104

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., … Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature, 498*(7452), E1–E2. https://doi.org/10.1038/nature12194

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. Retrieved from http://www.jstor.org/stable/2822435

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Turmunkh, U., van den Assem, M. J., & van Dolder, D. (n.d.). Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show. *Management Science*, (November 2018). https://doi.org/10.2139/ssrn.2919331

Turnbull, O. H., Bowman, C. H., Shanker, S., & Davies, J. L. (2014). Emotion-based learning: insights from the Iowa Gambling Task. *Frontiers in Psychology*, *5*, 1–11. https://doi.org/10.3389/fpsyg.2014.00162

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002

van den Assem, M. J., van Dolder, D., & Thaler, R. H. (2012). Split or Steal? Cooperative Behavior When the Stakes Are Large. *Management Science*, *58*(1), 2–20. https://doi.org/10.1287/mnsc.1110.1413

Van Lange, P. A. M., & Joireman, J. A. (2008). How We Can Promote Behavior That Serves All of Us in the Future. *Social Issues and Policy Review*, *2*(1), 127–157. https://doi.org/10.1111/j.1751-2409.2008.00013.x

Van Zant, A. B., & Kray, L. J. (2014). "I can't lie to your face": Minimal face-to-face interaction promotes honesty. *Journal of Experimental Social Psychology*, *55*, 234–238. https://doi.org/10.1016/j.jesp.2014.07.014

van Zuijen, T. L., Simoens, V. L., Paavilainen, P., Näätänen, R., & Tervaniemi, M. (2006). Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study. *Journal of Cognitive Neuroscience*, *18*(8), 1292–1303. https://doi.org/10.1162/jocn.2006.18.8.1292

Vaughan-Evans, A., Trefor, R., Jones, L., Lynch, P., Jones, M. W., & Thierry, G. (2016). Implicit Detection of Poetic Harmony by the Naïve Brain. *Frontiers in Psychology*, *7*, 1859. https://doi.org/10.3389/fpsyg.2016.01859

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press. https://doi.org/10.2307/2019327

Vyrastekova, J., & Van Soest, D. (2008). On the (in)effectiveness of rewards in sustaining cooperation. *Experimental Economics*, *11*, 53–65. https://doi.org/10.1007/s10683-006-9153-x

Walker, J. M., & Halloran, M. A. (2004). Rewards and Sanctions and the Provision of Public Goods in One-Shot Settings. *Experimental Economics*, *7*(3), 235–247. https://doi.org/10.1023/B:EXEC.0000040559.08652.51

Wang, C. S., Galinsky, A. D., & Murnighan, J. K. (2009). Bad Drives Psychological Reactions, but Good Propels Behavior Responses to Honesty and Deception. *Psychological Science*, *20*(5), 634–644. Retrieved from http://journals.sagepub.com/doi/pdf/10.1111/j.1467-9280.2009.02344.x

Warner, R. M., Kenny, D. a., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*(10), 1742–1757. https://doi.org/10.1037/0022-3514.37.10.1742

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology*, *12*(12), 1012–1015. https://doi.org/https://doi.org/10.1016/S0960-9822(02)00890-4

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852. https://doi.org/10.1126/science.288.5467.850

Welsh, N. A. (2004). Fairness: Perceptions of Fairness in Negotiation. *Marquette Law Review*, *87*(4). Retrieved from http://scholarship.law.marquette.edu/mulr

Wilkinson, L., & Shanks, D. R. (2004). Intentional Control and Implicit Sequence Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 354–369. https://doi.org/10.1037/0278-7393.30.2.354

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1047–1060. https://doi.org/10.1037/0278-7393.15.6.1047

Wu, Y. J., & Thierry, G. (2010). Chinese-English Bilinguals Reading English Hear Chinese. *Journal of Neuroscience*, *30*(22), 7646–7651. https://doi.org/10.1523/JNEUROSCI.1602-10.2010

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), 7398–7401. https://doi.org/10.1073/pnas.0502399102

Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(Supplement 2), 10424–10429. https://doi.org/10.1073/pnas.1301210110

Zajonc, R. B. (1965). Social facilitation. *Science, New Series*, *149*(3681), 269–274. Retrieved from http://www2.psych.ubc.ca/~schaller/Psyc591Readings/Zajonc1965.pdf

Zajonc, R. B., Heingartner, A., & Herman, E. M. (1969). Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology*, *13*(2), 83–92. https://doi.org/10.1037/h0028063

Zelmer, J. (2003). Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics*, *6*, 299–310.

Appendices

Appendix A

Public Goods Game Standardised Instructions

Note. The instructions below were used for the face-to-face version of the punishment/reward games. We then adapted these instructions for use in the anonymised punisment/reward games by eliminating the step-by-step instructions for the experimenter.

Introduction

- Welcome to the experiment
- This study examines individual and group investment behaviour. The money that you earn in this experiment you will be able to keep at the end of the task. At the end, the money that you have in your individual 'bank' will be changed up from pennies into larger cash denominations.
- If you follow the instructions, you can earn a fair sum of money.
- You will be playing an investment game where you will be given an endowment of 10p before each round.
- You will then choose how much of this endowment to contribute to a 'group pot'. This can be any proportion of your endowment, ranging from 0p to the entire 10p. Do not invest more than 10p on any round, otherwise the remaining amount exceeding 10p will be confiscated and not returned to you.
- What you earn will depend on how much the group invests. The more the group invests, the more each player earns. Always bear in mind the amount that you don't invest in relation to the amount you get back from the group. For example if you invest all 10p, if the return from the group pot is 9p, you have lost 1p. Similarly, if you contribute none of you 10p and even if the group return is 2p the total from that round that you will get is 12p.
- Contributions will be counted up, multiplied by 1.6 and split four ways. If the total is not divisible by four, it will be rounded up to a sum that is so that every person will earn the same amount. For example, if you decided to invest nothing into the group pot, but the other 3 members invested a total of 20p, you would get a return of 8p each. Every player will get the same return from the group pot, regardless of their original investment.
- You will be told the total of the group pot and how much your return is, before the money is returned to you.
- You will keep your return and any amount that you do not invest, in your individual 'bank' (the larger box in front of you).
- You must not share information about your contributions or engage in any negotiations with other players.
- Are there any questions?
- Please turn your screens so that other players cannot see what you are doing.

Trial procedure (practise rounds)

- This is round X
- You will now receive your endowment of 10p.

- Please choose how much of your endowment to invest (do not exceed 10p).
- Place the amount you want to invest into the box with the lid and hand it to the experimenter. Please keep the rest of your endowment on the mat in front of you.
- The banker will now tally the investments.
- Your group pot was X, which was multiplied by 1.6 and then divided by 4, giving an individual return of Y.
- Place your return and your leftover endowment into your bank.

Punishment/Reward introduction

- In the next rounds, you will now be able to punish/reward other players.
- You will now receive a set of punishment/reward tokens. Each of you will receive the same number of tokens. Each token allows you to punish/reward one player, once. You may only punish/reward one player per round.
- If you choose to punish/reward someone on any particular round, you will place a punishment/reward token into your box together with your contribution to the group pot.
- The banker will then give the punishment/reward to the player who has made the lowest/highest contribution in that round. If two or more players are tied for the lowest/highest return, the punishment/reward will be determined at random. It is therefore possible to punish/reward yourself if you are the lowest/highest contributor.
- [Monetary punishment/reward] You will know if you receive a punishment/reward because the banker will place a black/gold coin into the box of the punished/rewarded player. The player's return from the group pot will also be reduced/increased by 50%. So for example, if the group return is 4, the punished/rewarded player will receive 2/6. [Monetary punishment only] If the return is not divisible by 2, it will be rounded down to the nearest number. So, if the return is 5, a punished player's actual return would be 2.
- [Social punishment/reward] You will know if you receive a punishment/reward because the experimenter will place a black/gold coin into the pouch on the top of the screen in front of you. This will be publically visible to other players for one round, after that the token will remain in your bank.
- [Monetary + Social punishment/reward] You will know if you receive a punishment/reward because the experimenter will place a black/gold coin into the pouch on the top of the screen in front of you. This will be publically visible to other players for one round, after that the token will remain in your bank. Also, your return from the group pot will be reduced/increased by 50%. So for example, if the group return is 4, the punished/rewarded player will receive 2/6. [Monetary + Social punishment only] If the return is not divisible by 2, it will be rounded down to the nearest number. So, if the return is 5, a punished player's actual return would be 2.
- The identity of the punisher/rewarder and the punished/rewarded must remain anonymous. If more than one person chooses to punish/reward on any round, only one player will be punished/rewarded. The other punishment/reward will be saved for a round where no one has chosen to punish/reward.
- Here are your punishment/reward chips.
- Do you have any questions?

Trial procedure (Monetary punishment/reward rounds)

- You will now receive your endowment of 10p

- You may now choose how much of your endowment to invest into the group pot and whether you will punish/reward other players in this round.
- Place your contribution, and if you have chosen to punish/reward, your punishment/reward token, into your box and place the box in the middle of the table.
- Please do not tell anyone if you have given a punishment/reward on this round.
- The banker will now count up your contributions and determine punishments/rewards.
- The group pot was X, which was multiplied by 1.6 and then divided by 4, giving an individual return of Y.
- If you have been punished/rewarded, a black/gold coin will appear in your box and your return will be cut in half/doubled. Please do not tell anyone if you have been punished/rewarded or if you gave out the punishment/reward.
- [Reward only] If you receive a gold coin, please do not eat it until the end of the game.
- Place your return into your bank.

Trial procedure (Social punishment/reward rounds)

- You will now receive your endowment of 10p
- You may now choose how much of your endowment to invest into the group pot and whether you will punish/reward other players in this round.
- Place your contribution, and if you have chosen to punish/reward, your punishment/reward token, into your box and place the box in the middle of the table.
- Please do not tell anyone if you have given a punishment/reward on this round.
- The banker will now count up your contributions and determine punishments/rewards.
- The group pot was X, which was multiplied by 1.6 and then divided by 4, giving an individual return of Y.
- Player blue/red/yellow/green receives a black/gold coin.
- Place your return into your bank.
- [Reward only] Please do not eat your gold coin(s) until the end of the game.

[EXPERIMENTER: rewards should be visible for one round of the game. After that, a punished/rewarded individual should remove the gold coin and place it behind the screen (in their bank) until the end of the game].

Trial procedure (Monetary + Social punishment/reward rounds)

- You will now receive your endowment of 10p
- You may now choose how much of your endowment to invest into the group pot and whether you will punish/reward other players in this round.
- Place your contribution, and if you have chosen to punish/reward, your punishment/reward token, into your box and place the box in the middle of the table.
- Please do not tell anyone if you have given a punishment/reward on this round.
- The banker will now count up your contributions and determine punishments/rewards.
- The group pot was X, which was multiplied by 1.6 and then divided by 4, giving an individual return of Y.
- Player blue/red/yellow/green receives a black/gold coin

- Place your return into your bank.
- [Reward only] Please do not eat your gold coin(s) until the game has ended.

[EXPERIMENTER: punishments/rewards should be visible for one round of the game. After that, a punished/rewarded individual should remove the gold coin and place it behind the screen (in their bank) until the end of the game].

End of the game

- You have reached the end of the experiment.
- I will give you each a packet of questionnaires to complete. While you complete the questionnaires, the banker will exchange your pennies for larger coin denominations. These will be returned to you at the end of the session.
- Here are your questionnaires.

Debriefing

- This experiment was about the effect of different types of reward on investment behaviour.
- More specifically, it concerned how punishment alters the likelihood of "free-riding" or investing nothing into the group pot.
- So, it was looking at whether punishments cause people to act in a less self-interested way and how the opportunity for punishment, anonymity and reputation has a bearing on this behaviour.
- Do you have any questions about the experiment or the game?
- You earned £X.XX today. I'll need you to sign a receipt saying you received this amount.

Appendix B (i)

Comprehension quiz

1. What game strategy will allow you to earn the most money in the game if everyone else contributes **almost all** their money in the game?

    a) Contributing almost all my endowment on each trial

    b) Contributing half my endowment on each trial

    c) Contributing almost none of my endowment on each trial

2. What game strategy will allow you to earn the most money in the game if everyone else contributes **almost none** of their money in the game?

    a) Contributing almost all my endowment on each trial

    b) Contributing half my endowment on each trial

    c) Contributing almost none of my endowment on each trial

3. How likely is it that all the other players will contribute more than half their endowment in the game?

    a) Not at all likely

    b) Somewhat likely

    c) Very likely

4. Which of the following outcomes will occur if you contribute almost all of your money and the rest of the players contribute almost none of theirs (circle all that apply)?

    a) The other players will earn more money than me.

    b) I will earn more money than the other players.

    c) My contribution will benefit the other players more than it will benefit me.

    d) I will lose money in the long run because I will keep less of my endowment.

Appendix B (ii)

Punishment game comprehension scoring & analysis

On this quiz, we asked two types of question – the first assessed a player's comprehension of the game rules and the second assessed their belief in how other players may behave.

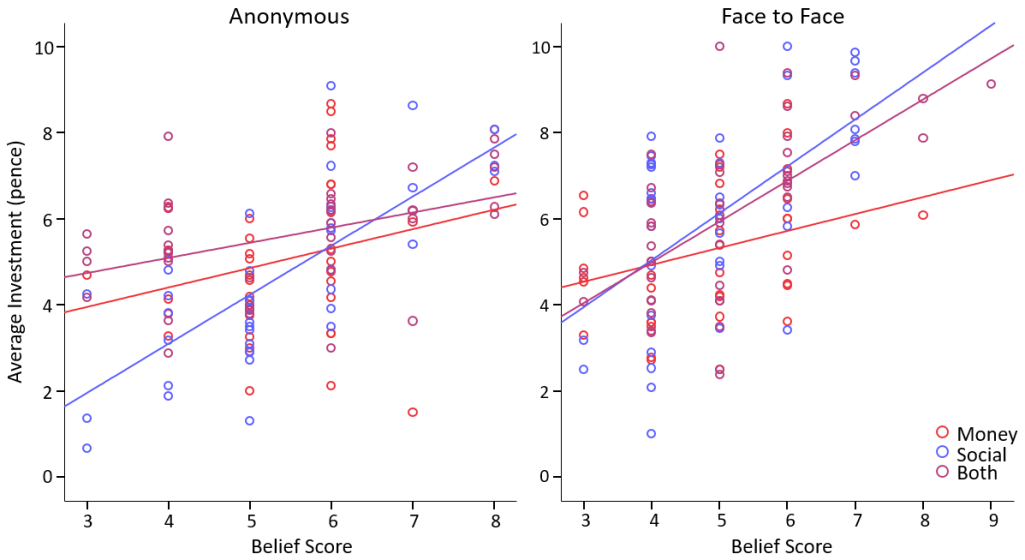
Comprehension question scoring: on Question 4, there were multiple correct answers and we asked participants to select as many as they deemed appropriate. In hindsight, this was probably not the most effective way to ask this question, given that this is an atypical format to a multiple choice question for students at Bangor University. Nonetheless, we awarded one point for each of the correct options of a), c) and d). We deemed that if a player had selected option b), this demonstrated a definite lack of understanding, and so were penalised by being docked one point (-1). Therefore the range of 'Comprehension' points available was: -1 to 3.

Belief question scoring: We designed Questions 1, 2 & 3 to assess players' belief in how others might play the game. These questions did not have 'correct' answers per se. Rather, they were designed to assess each player's individual preferences. For example, on Question 1, a player may believe that when all other players are investing highly, they will earn the most when they also invest highly. However, a player may also believe that if other players are investing highly, that they can earn the most by investing very little (retaining most of the endowment plus allowing others to subsidize their returns). Therefore, we decided to score this question on a scale of 1-3, with higher scores indicating that a player has chosen a more a cooperative belief. Thus, if a player selected option a) they received 3 points; b) received 2 points (as a neutral option); c) received 1 point. Question 2 was scored in the same way. On Question 3, those who selected option a) indicated a belief that other players were non-cooperative and so this option was assigned 1 point; b) received 2 points; c) received 3. Therefore, the range of 'Belief' points available was: 3 to 9.

Player comprehension scores did not differ across punishment condition in the anonymous game, $F(2,117)=0.41$, $p=.667$, $\eta_p^2=.01$, or in the face-to-face game, $F(2,121)=0.60$, $p=.549$, $\eta_p^2=.01$. Thus players had a similar grasp of the game across all conditions. Players also started out with similar belief levels, regardless of punishment condition, both in the anonymous context – $F(2,117)=0.06$, $p=.940$, $\eta_p^2=.01$; and in the face-to-face context - $F(2,121)=0.87$, $p=.420$, $\eta_p^2=.01$.

In terms of how comprehension scores related to investment, we found no relationship between understanding and investment behaviour. In both the anonymous and face-to-face games, there was no relationship between average player investment and comprehension score (p-values $\geq .250$).

However, in terms of belief scores, we found that those players who believed that others were more cooperative actually cooperated more themselves, by investing more on average. This was the case in both the anonymous game, $r(120)= .48$, $p<.001$, and in the face-to-face game, $r(124)= .53$, $p<.001$ (see figure below). This result fits with previous research that finds a positive relationship between the belief that others are fair/helpful and their own contribution to the Public Good (Gächter, Herrmann, & Thöni, 2004).



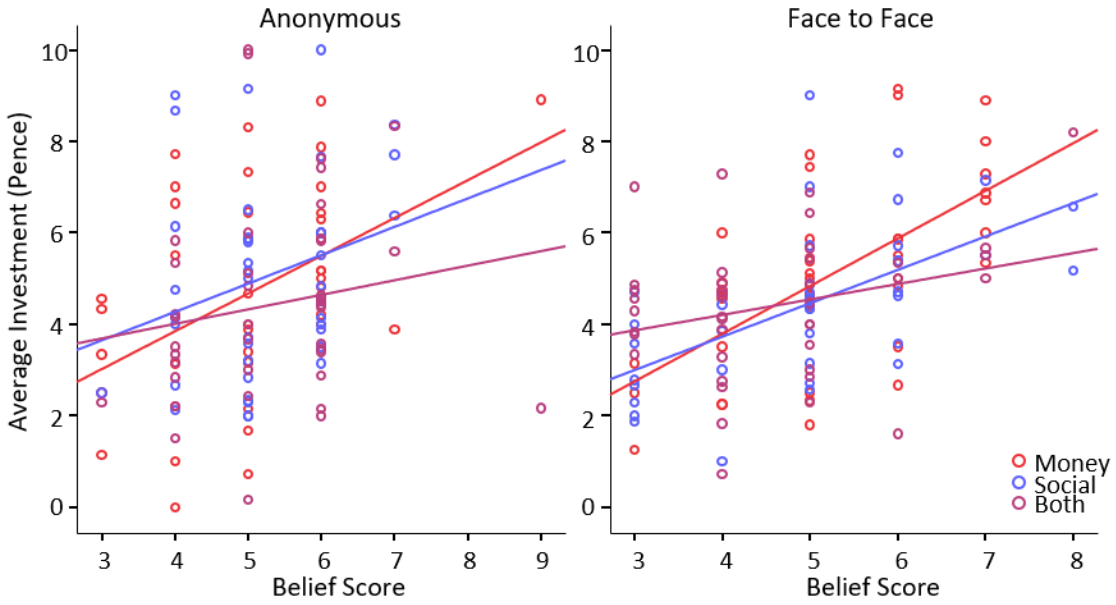
	Punishment Condition	Mean comprehension (*SD*)	Mean belief (*SD*)
Anonymous	Money	1.65 *(1.11)*	5.40 *(1.07)*
	Social	1.50 *(1.12)*	5.33 *(1.27)*
	Both	1.73 *(1.14)*	5.43 *(1.55)*
Face-to-face	Money	1.63 *(0.94)*	4.80 *(1.19)*
	Social	1.50 *(0.95)*	4.98 *(1.19)*
	Both	1.39 *(1.05)*	5.16 *(1.3)*

Appendix B (iii)

Reward game comprehension scoring & analysis

Player comprehension scores did not differ across punishment condition in the anonymous game, $F(2,117)=0.84$, $p=.062$, $\eta_p^2=.05$, or in the face-to-face game, $F(2,116)=0.48$, $p=.623$, $\eta_p^2=.01$. Players also started out with similar belief levels, regardless of punishment condition, both in the anonymous context – $F(2,117)=0.62$, $p=.539$, $\eta_p^2=.01$; and in the face-to-face context - $F(2,116)=2.66$, $p=.075$, $\eta_p^2=.04$.

We found no relationship between understanding and investment behaviour. In both the anonymous and face to face games, there was no relationship between average player investment and comprehension score (p-values $\geq .228$). However, there was a significant relationship between average investment and belief score, in both the anonymous game, $r(120)=.30$, $p=.001$, and in the face to face game, $r(119)= .51$, $p<.001$ (see graphs below). This would suggest that the more that players believe that others will play cooperatively, the more that they themselves invest on average.



	Reward condition	Mean comprehension (*SD*)		Mean belief (*SD*)	
Anonymous	Money	2.11	*(0.93)*	5.07	*(1.19)*
	Social	1.70	*(1.08)*	5.20	*(0.93)*
	Both	1.63	*(1.18)*	5.30	*(1.12)*
Face-to-face	Money	1.50	*(0.92)*	5.23	*(1.17)*
	Social	1.53	*(0.92)*	4.93	*(1.37)*
	Both	1.33	*(0.82)*	4.58	*(1.24)*

Appendix C

Standard game strategy reminder

- What you must consider is how much you put into the group in relation to how much you are getting back.
- For example, if you invest 5p of your 10p endowment, and you get 4p returned from the group, you have lost a penny. Your total on that round would be 9p, which is less than your original endowment.
- You don't want to be losing money at any point during the game, so be prepared to readjust your strategy.
- Always be asking yourself whether you are winning or losing money on each round.
- Also consider what other people may be investing and how that may affect your own return. You may end up subsidising other players' payoffs if you are investing more than others are.

Appendix D

Video rating questionnaire

Response scale 1:

Not at all	Very little	Somewhat	Quite a bit	A great deal
1	2	3	4	5

1. How engaged with the task did the players seem to be?
2. Did there seem to be any tension between players?
3. How much laughter and/or smiling was evident during the interaction?
4. How talkative did the group seem to be?
5. How smooth and coordinated did the interaction seem?
6. If you were playing a game in which there were two-person teams and your goal was to win the game, how much would you like to be on the same team with:
    a. The Red player
    b. The Blue player
    c. The Green player
    d. The Yellow player

Response scale 2:

Very Negative	Negative	Neutral	Positive	Very Positive
1	2	3	4	5

7. What was the general mood of the group?

Appendix E

Average investment 3-way ANOVA

To examine average investment across all conditions in both experiments, we conducted a 3-way between groups ANOVA. Here, we entered Experiment (Punishment/Reward), Interaction Condition (Anonymous/Face-to-face) and Incentive Type (Money/Social/Both) as terms in the model. Results of this analysis showed a main effect of Experiment - $F(1,480)=44.76$, $p<.001$, $\eta_p^2=.09$ - investments in the Punishment experiment were on average larger ($M=6.59$, $SD=1.90$) than in the Reward experiment ($M=5.39$, $SD=2.16$). There was also a difference between the Interaction Conditions across both experiments, $F(1,480)=7.16$, $p=.008$, $\eta_p^2=.01$, with investment in the face-to-face condition being larger on average ($M=6.25$, $SD=2.21$) than that of the anonymous condition ($M=5.73$, $SD=2.01$) .

Additionally, there was a main effect of incentive type $F(2,480)=3.22$, $p=.041$, $\eta_p^2=.01$, with players investing less in the Monetary ($M=5.94$, $SD=2.03$) and Social ($M=5.73$, $SD=2.34$) conditions compared to the Combined condition ($M=6.29$, $SD=1.96$). *However*, the presence of a significant Experiment by Incentive Type interaction, $F(1,480)=4.96$, $p=.007$, $\eta_p^2=.02$, suggests that this effect differs depending on the experimental setting (punishment/reward). This interaction is thus broken down further within comparative models in Chapters 2 & 3.