

Experimental Essays on Morality and Perception

Kevin P. Grubiak



Thesis submitted for the degree of
Doctor of Philosophy
in Economics

University of East Anglia
School of Economics

September 2020

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

This PhD thesis is a collection of three independent essays employing experimental methods to investigate the links between moral behaviour and perception. Chapter 1 explores the role of image concerns in promise keeping. In our baseline treatments, we use double dictator games which embed and vary opportunities for subjects to hide their selfishness through self- and other-deception. Adding opportunities for promise exchange, our data is consistent with social-image concerns as one motivator of promise keeping. We find no evidence of subjects engaging in self-deception to evade their promise-induced commitments. Chapter 2 explores motivated reasoning in a context where third-party bystanders can prevent future norm transgressions. For this purpose, we introduce the Third-Party Protection Game. In this game, a third-party player can invest own resources to protect a passive player's endowment from being appropriated by a dictator. The game features uncertainty regarding the degree of protection needed. We hypothesise that third-parties will report conveniently biased, i.e., less cynical beliefs about dictators the costlier it is to protect. Our data only provides moderate support. What we do find however is that third-parties more generally and irrespective of the assigned cost overestimate dictator generosity. Chapter 3 introduces the Costless Sharing Game (CSG). In this game, a sharer first earns a resource by completing a task and is then offered the opportunity to share the resource at no personal cost with a recipient. We use the CSG to consider how sharing depends on moral reasoning based on entitlement and desert ("intrinsic moral motivation") and on whether the context of the sharer's decision is known by the recipient ("extrinsic social motivation"). We observe very little reluctance to share. Interestingly, we also find mild evidence of a treatment interaction which suggests less sharing when neither intrinsic moral nor extrinsic social arguments for sharing are present.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Contents

List of Figures	vi
List of Tables	vii
Acknowledgements	viii
Introduction	1
Chapter 1: Exploring Image Motivation in Promise Keeping	4
1.1 Introduction	5
1.2 Related Literature	6
1.2.1 Promise Keeping	6
1.2.2 Social-Image Concerns	8
1.2.3 Self-Image Concerns	9
1.3 The Experiment	10
1.3.1 Design	10
1.3.2 Hypotheses	16
1.3.2.1 Social-Image Concerns	17
1.3.2.2 Self-Image Concerns	18
1.3.3 Procedures	19
1.4 Results	19
1.4.1 Communication Contents	20
1.4.2 Communication Effects	20
1.4.2.1 Social-Image Effects	21
1.4.2.2 Self-Image Effects	25
1.5 Discussion	28
1.6 Conclusion	30
Appendix for Chapter 1	31
1.A Supplementary Data	31
1.A.1 Cut-offs and Task Performance	31
1.A.2 Regression Results	32

1.B	Instructions and Screens	33
1.B.1	Main Treatment Instructions	33
1.B.2	Practice Stage Screens	37
1.B.3	Control Questions	45
1.B.4	Control Treatment Instructions	47
1.B.5	Revelation of Cut-off Details	50
 Chapter 2: Third-Party Intervention and Perception Manipulation		51
2.1	Introduction	52
2.2	Related Literature	53
2.2.1	Third-Party Intervention	53
2.2.2	Perception Manipulation	55
2.3	Experiment 1	57
2.3.1	Design	57
2.3.2	Hypotheses	60
2.3.3	Procedures	61
2.3.4	Results	62
2.3.4.1	Protection Behaviour	62
2.3.4.2	Beliefs about the Claim	64
2.4	Experiment 2	67
2.4.1	Design	67
2.4.2	Procedures	68
2.4.3	Results	68
2.5	General Discussion	70
2.6	Conclusion	71
	Appendix for Chapter 2	72
2.A	Sequential Game Comparison	72
2.B	Instructions and Screens	72
2.B.1	Experiment 1 Instructions	72
2.B.2	Example Round Screens	77
2.B.3	Control Questions	85
2.B.4	Experiment 2 Instructions	87
2.B.5	Strategy Method Example Screens	92

Chapter 3: Costless Sharing, Moral Entitlements and Perception	93
3.1 Introduction	94
3.2 Related Literature	95
3.2.1 Dictator Game Giving	95
3.2.2 Costless Sharing	96
3.3 The Experiment	97
3.3.1 Design	97
3.3.2 Hypotheses	99
3.3.2.1 Intrinsic Moral Motivation	99
3.3.2.2 Extrinsic Social Motivation	99
3.3.3 Procedures	99
3.4 Results	100
3.4.1 Main Results	100
3.4.2 Supplementary Results	102
3.5 Discussion	104
3.6 Conclusion	105
Appendix for Chapter 3	106
3.A Instructions and Screens	106
Bibliography	113

List of Figures

Figure 1.1: Sequence of Stages in the Experiment	11
Figure 1.2: Calibrated Cut-off Distribution	13
Figure 1.3: Proportions of Generous Choices between Treatments	23
Figure 1.4: Cumulative Distribution Functions of Success Times	26
Figure 1.5: Average Times Taken per Task	27
Figure 2.1: Player C's Decision Screen	58
Figure 2.2: Belief Elicitation Screen	60
Figure 2.3: Average Protection Behaviour by Treatment and Role Order	62
Figure 2.4: Beliefs and Protection Behaviour by Treatment	64
Figure 2.5: Average Beliefs by Treatment and Role Order	65
Figure 2.6: Distributions of Beliefs and Empirical Claims	66
Figure 2.7: Player C's Decision Screen for Scenario 1	68
Figure 3.1: Distributions of Tokens Shared	101
Figure 3.2: Distributions of Tokens Shared (Reduced Sample)	104

List of Tables

Table 1.1:	Factorial Treatment Design	10
Table 1.2:	Belief Elicitation	15
Table 1.3:	Overview of Message Profiles by Treatment	20
Table 1.4:	Summary Statistics	22
Table 1.5:	Success Times and Accuracy in the Matrix Task	25
Table 1.6:	Random Effects Panel Model Estimations	28
Table A.1:	Cut-offs and Task Performance	31
Table A.2:	Regression Results	32
Table 2.1:	Summary Statistics for Experiment 1	63
Table 2.2:	Summary Statistics for Experiment 2	69
Table 2.3:	Random Effects Tobit Model Estimations	69
Table 3.1:	Factorial Treatment Design	98
Table 3.2:	Summary Statistics of Tokens Shared	100
Table 3.3:	Classification of Written Explanations	102
Table 3.4:	Control Question Attempts	103
Table 3.5:	Summary Statistics of Tokens Shared (Reduced Sample) . . .	103

Acknowledgements

I am indebted to many people who have supported me during this challenging yet exciting and very rewarding process of writing this thesis. First of all, I want to thank my advisors Prof. Robert Sugden, Dr. Anders Poulsen and Dr. Odile Poulsen for giving me the freedom to pursue my own research interests and for the numerous discussions we had over the years which improved my work and made me a better thinker. Bob, you are one of the smartest and kindest people I know; thank you for your guidance and encouragement. Anders, thank you for your advice, encouragement and for involving me in many exciting research projects. Odile, thank you for your interest and the insightful comments you made on early drafts of my work; chapter 3 of this thesis is dedicated to you and your daughter Eva.

I particularly want to thank my friends in the PGR community and beyond for being there for me when I needed it and for having made my stay in the U.K. such an unforgettable experience. I share memories with you that will last for a lifetime.

The unconditional support of my family helped me to keep going and to stay focused. I thank my mother and my brother for always having my back, I can't express how much this means to me.

Finally, I want to thank the School of Economics as well as the attendees of numerous seminars and conferences for their feedback on my work. I am also grateful to Prof. Christoph Vanberg and Dr. Amrish Patel for kindly accepting to serve as discussants of this thesis.

Kevin Grubiak
University of East Anglia
September 2020

Introduction

Insights from behavioural and experimental economics have enriched the way we think about humans, their objectives and the process of decision making. As a graduate student, I got excited reading about human psychology and the behavioural relevance of social preferences which stood in sharp contrast to a model of behaviour based on rationality and pure self-regard. Subjects in carefully designed experiments were often times altruistic and seemed to care about notions of fairness, reciprocity, intentions, or norm compliance. A recent literature has challenged these findings by demonstrating that many subjects are reluctantly pro-social, i.e. they pursue moral objectives for extrinsic reasons such as pride, guilt and shame, but cease to do so when such incentives are removed. What motivates behaviour in such instances is not morality per se, but the desire of being *perceived* as a moral person. A lot of the evidence on perception concerns stems from dictator game studies which focus on testing the robustness of a distributional fairness norm. It is important not to stop here however, and to assess the extent to which perception concerns play a role in explaining the strength of other norms frequently found to be behaviourally important; this is the objective of my thesis. I present an analysis of three novel experiments, each dealing with the relevance of perception concerns applied to a different morally demanding context.

I use perception concerns as an umbrella term subsuming two distinct factors that may matter for behaviour: social-image concerns and self-image concerns. Whereas in the former case, behaviour is expected to depend on how others perceive a given action, in the latter case what matters is how actions reflect on a decision-maker's self-perception. The standard approach of manipulating social-image concerns in the lab involves comparing treatments where other subjects are fully informed about a decision maker's actions and the associated consequences with treatments where no such information is provided. Self-image concerns, on the other hand, are distinct in that they don't require observers to be present. One could think of the decision maker as being his or her own observer and internal assessor of behaviour. Under this perspective, seemingly moral actions may be the result of subjects avoiding having to send a signal about their type that would threaten their self-image. One way of mitigating such concerns is to engage in motivated reasoning or self-deception which can be tested for in the lab.

In chapter 1, I investigate the role of image concerns in promise keeping using a lab experiment. The experiment I designed allows to test for both self-image and social-image concerns in a unified framework. In my baseline treatment, I combine a

double dictator game with an effort task that has to be completed successfully for the dictator to be able to decide between a self-regarding and a generous allocation. If a subject fails to complete the task, the computer implements a random allocation between the two alternative allocations. Importantly, the task is subject to a random cut off mechanism which may prevent subjects from completing the task on time in which case the decision is delegated to the computer. Recipients do not learn whether a decision was made by the dictator or the computer. This design feature is a modification of the “plausible deniability” mechanism introduced by Dana, Weber and Kuang (2007) and allows me to measure whether subjects work reluctantly, or procrastinate, on the task to delegate their decision to the computer. Such procrastination (which is measured in relation to a control treatment where we remove incentives for procrastination) is in line with a self-deceptive strategy according to which subjects *pretend* to be interested in solving the task, but do so half-heartedly in the hope of obtaining the self-regarding allocation through the computer. Promise keeping enters the picture by adding a pre-play communication stage to the baseline treatment which allows subjects to exchange pre-formulated promises about their intent to solve the task and to choose the generous allocation. A comparison of response times and task accuracy between treatments with and without communication is then indicative of whether or not subjects were reluctant to live up to their promise-induced commitments. To shed light on the social-image dimension, I also add treatments where recipients are able to infer whether an allocation came about by the dictator or the computer. This completes my 2x2 design. My results show that social-image concerns matter for promise keeping. Still, I find significant promise keeping rates even when actions are deniable and no evidence of subjects engaging in self-deception to evade their promise-induced commitments. This could be interpreted as corroborating evidence of the strength of promises.

In chapter 2, I investigate the relevance of motivated reasoning in a context where third-parties or bystanders can intervene to prevent future norm transgressions. For this purpose, I introduce the Third-Party Protection Game and implement it in the lab. In this game, a third-party player can invest own resources to protect another passive player’s endowment from being appropriated by a dictator. Importantly, the third-party has to decide on the level of protection ex-ante, i.e. before the required level of protection – as determined by the dictator’s decision – is revealed. The question I pose is: will third-parties exploit the inherent uncertainty about the dictator’s behaviour in a self-serving way and convince themselves that a norm transgression is unlikely or less severe? I elicit third-parties’ beliefs about dictator behaviour and provide significant incentives for accurate beliefs. To check for evidence of distorted beliefs, I follow a strategy introduced by Di Tella et al. (2015) whereby a decision maker (in this case, the third-party) is privately informed

of their assignment to one of two (protection) cost conditions: low cost vs. high cost. Since the cost assignment is private knowledge to third-parties and dictators are kept uninformed, third-party beliefs about dictator behaviour should not differ systematically between treatments. In contrast, if third-parties entertain beliefs which are motivated by a desire to avoid costly protection, these beliefs may reflect less cynicism the costlier it is for third-parties to protect. What I find are differences in beliefs between the two cost conditions in the direction hypothesised by motivated beliefs; these differences however only reach mild significance. I however do observe that beliefs matter and that third-parties, more generally and irrespective of the assigned cost condition, expect dictators to be less selfish than they really are. This suggests an ability of policy makers to affect behaviour in the field by disseminating more accurate information about the severity of norm transgressions.

In chapter 3, I present the results of an online experiment jointly designed with Anders Poulsen, Mengjie Wang and Jiwei Zheng. The online implementation of the experiment was a consequence of the social distancing requirements associated with the Covid-19 pandemic which made us deviate from our initial intention to run the experiment in the physical lab. For an excellent discussion of the pros and cons of running experiments in the physical lab versus online we refer the reader to an article by Horton, Rand and Zeckhauser (2011). We couldn't identify any relevant limitations from moving our experiment online. On the contrary, we liked the online version on the grounds that it made the experimenter-subject anonymity of our experiment more credible. The experiment introduced the Costless Sharing Game where a sharer first earns an endowment by completing an effort task and is then offered the opportunity to share the resource at no personal cost with another person, the recipient. We think that the empirical relevance of costless sharing is significant; examples include emailing presentation slides, sharing documents, and more generally sharing valuable information, knowledge, and advice with someone else. To our knowledge, very little is known about people's willingness to share when resources are excludable but non-rival. We use the Costless Sharing Game to consider how the amount shared depends on moral reasoning based on entitlement and desert ("intrinsic moral motivation") and on whether the context of the decision of the sharer is known by the recipient ("extrinsic social motivation"). Similar to Cappelen et al. (2017), we manipulate the first channel by comparing treatments where recipients are passive with treatments where recipients had to successfully complete the same task that gave rise to the sharer's resource. We manipulate the second channel by varying the information that recipients receive about the context and decision of the sharer. This completes our 2x2 design. Our results suggest very little reluctance to share. Interestingly, we also find mild evidence of an interaction between our treatment conditions which indicates less sharing when neither intrinsic moral nor extrinsic social arguments for sharing are present.

Chapter 1:

Exploring Image Motivation in Promise Keeping*

*I would like to thank Robert Sugden, Anders Poulsen, and Odile Poulsen for financial support and helpful guidance. I would also like to thank David Hugh-Jones for serving as a discussant at the design stage of the experiment and Joël van der Weele and Kiryl Khalmetski for useful comments. Finally I would like to thank the audiences of the 2018 CCC (CBESS-CEDEX-CREED) meeting, the 14th TIBER Symposium on Psychology and Economics, and the 2019 European meeting of the Economic Science Association for their feedback.

1.1 Introduction

Trust plays an important role in many economic interactions. It is a prerequisite for interactions where legal contracts are not enforceable or simply too expensive to implement. Moreover, trust can provide substantial efficiency gains, for instance, by speeding up the process of decision making. Despite its potential benefits, however, trust carries the risk of betrayal.

Yet, abundant evidence documents that people are far more trustworthy than the standard economic model resting on the assumption of pure self-interest would assert. Prominent explanations relate to intrinsic preferences for concepts like fairness, equality, or reciprocity. But also factors like the ability to talk and exchange promises have widely been observed to increase trust and trustworthiness. The inclination to keep a promise can theoretically and empirically be accounted for by the *commitment-based* (Vanberg, 2008) as well as the *expectations-based* (Charness and Dufwenberg, 2006) explanations. According to the former, people keep their promises because they have an intrinsic preference for keeping their word. According to the latter, promises are kept because they induce a shift in promisee expectations and, thus, higher experienced guilt by the promise maker. Although these theories are not mutually exclusive, follow-up research has used ever more sophisticated experimental protocols in an attempt to cleanly distinguish between these two motivations of promise keeping (e.g., Vanberg, 2008; Schwartz, Spires and Young, 2019; Bhattacharya and Sengupta, 2016; Ederer and Stremitzer, 2017; Ismayilov and Potters, 2016; Mischkowski, Stone and Stremitzer, 2019; Di Bartolomeo et al., 2019). Although guilt aversion appears to play a significant role, promises are frequently kept even when guilt is ruled out as an explanation. On balance, these studies provide remarkable support of both an intrinsic preference and guilt aversion in promise keeping.

In contrast to the cited literature, our study does not aim to assess the empirical relevance of these competing theories of promise keeping. Instead, in the current paper we explore the relevance of alternative and understudied reasons for honouring one's word, namely *image concerns*. We say that a decision maker is concerned about his or her image if he or she experiences a disutility from being *perceived* in a negative light either by other individuals (social-image concern) or by him- or herself (self-image concern). Although there is a well-established literature documenting that these types of concerns indeed affect decision making in a variety of morally demanding contexts (Gino, Norton and Weber (2016); see also sections 1.2.2 and 1.2.3 for a detailed review of the respective literatures), very little is known about its particular relevance in the domain of promise keeping.

Two studies which address the social-image hypothesis in promise keeping are Deck, Servátka and Tucker (2013) and Schütte and Thoma (2014). Both report

a null result. As acknowledged by the authors themselves however, their lack of an effect could be due to the high rates of cooperation that both observe in their baseline conditions, consequently “leaving little room for incremental improvement in cooperation” (Deck, Servátka and Tucker, 2013, p. 598). Our paper contributes to this strand of the literature by documenting a positive result in an experiment where such “ceiling effects” are minimised.

The second contribution of our paper lies in its test of self-image concerns in promise keeping. To the best of our knowledge, all studies on promise keeping make it perfectly transparent to the promisor that he or she is responsible for a broken promise. In reality however, people can often excuse their behaviour in ways which allow them to preserve their self-image e.g. by shifting responsibility for outcomes to external circumstances. It turns out that our test of self-image concerns yields a null result: we find no evidence of subjects engaging in self-deception to evade their promise-induced commitments which could be interpreted as corroborating evidence of the strength of promises.

The remainder of this paper is structured as follows. Section 1.2 reviews the related literatures in more detail. Section 1.3 elaborates on the experimental design, the hypotheses and the procedures of our experiment. In Section 1.4 we present the results. Section 1.5 contains a discussion. Section 1.6 concludes the analysis.

1.2 Related Literature

Our study connects two strands of the literature which, by and large, have only been considered in isolation from each other: the literatures on *promise keeping* and on *image concerns*. In this section, we review each respective literature and outline how a joint perspective could improve our understanding of the effectiveness of non-binding verbal commitments.

1.2.1 Promise Keeping

Although standard economic theory discards an influence of pre-play communication on behaviour, numerous studies have documented that communication, in particular the use of promises, can substantially increase cooperation. Unaccounted for by the standard approach, people may be averse to lying or dislike letting others down on what they promised them they would do, which may eventually render cheap talk *credible*.

In a seminal paper, Charness and Dufwenberg (2006) introduce a hidden-action trust game with pre-play communication and find that promises significantly increase cooperation. The cooperative strategy profile occurred 20% of the time without communication and 50% of the time with communication. The authors argue that

their results square well with a model of guilt aversion by which promises feed expectations which the promisor dislikes to violate (*expectations-based* explanation).

Yet, a popular alternative explanation of their results is that people may hold an intrinsic preference for keeping their word (*commitment-based* explanation). A series of papers have been dedicated to disentangling these two explanations of promise keeping. The first of which, Vanberg (2008), uses a variant of the hidden-action trust game where subjects are informed that there is a 50% chance that they will be re-matched to a different subject than the one they previously communicated with. Only the promisor is informed whether his or her partner was switched and he or she is allowed to inspect the message that this new partner has received earlier, before the switch occurred (hence, the promisor knows whether or not a promise was received). From the perspective of the promisee who is unaware whether or not a switch occurred, first-order beliefs about the promisor's trustworthiness should not differ across conditions. Anticipating this, the promisor's second-order belief and hence the guilt potentially experienced should not differ either. Holding second-order beliefs constant, Vanberg finds that a dictator's own promise affects behaviour whereas a promise that was made by others does not.¹ He argues that this result appears to be incompatible with the expectations-based explanation of promise keeping and lends support to the commitment-based explanation.

Ederer and Stremitzer (2017) claim that the lack of evidence of guilt aversion in promise keeping observed by Vanberg (2008) may result from the possibility that guilt is only experienced if one is directly responsible for inducing an increase in the expectations of a promisee. Recall that in Vanberg's experiment, the increase in expectations in the control condition is induced by *another* dictator's promise, while expectations are affected by the dictator's *own* promise in the main condition. The authors use an "unreliable random device" to generate exogenous variation in second-order beliefs and provide evidence of guilt aversion in promise keeping. However, as their design does not include an analogue to Vanberg's partner-switching mechanism, they cannot assess the strength of the expectations-based explanation relative to the strength of the commitment-based explanation.

In a unified framework, Di Bartolomeo et al. (2019) study an environment that allows for exogenous variation of *both* promises and expectations allowing them to test which channel is quantitatively more important. They essentially combine the earlier designs by Vanberg (2008) and Ederer and Stremitzer (2017). More precisely, they make the partner-switching probability in Vanberg's design a separate treatment variable that randomly takes the value *low* (25%) or *high* (75%) to generate exogenous variation in expectations. Their results suggest that promise keeping is *independent* of beliefs. Promise keeping rates are as high when beliefs are

¹Dictators who promised chose the cooperative outcome 73% of the time whereas those who did not promise chose it 52% of the time.

low (as induced by a high switch probability) as when beliefs are high (as induced by a low switch probability). Nonetheless, they observe an independent effect of higher expectations on cooperation as predicted by guilt aversion.

The overall picture documents that (i) the use of promises is a powerful way of increasing cooperation and efficiency and (ii) that its effect is mediated by *both* an intrinsic preference for promise keeping and guilt aversion. Yet, another motivation for promise keeping which has received little attention so far is *image* motivation.²

1.2.2 Social-Image Concerns

Social-image concerns relate to people's concerns over how their actions are perceived by others. A well-established literature has documented that choices depend on observability (see e.g. Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Ariely, Bracha and Meier, 2009; Bohnet and Frey, 1999; Bursztyn and Jensen, 2017; Chaudhuri, 2011; Dana, Cain and Dawes, 2006; Ekström, 2012; Fehr and Gächter, 2000; Rege, 2004; Rege and Telle, 2004; Soetevent, 2005; Tadelis, 2011). Altruistic behaviour in the well-known dictator game, for instance, has been shown to be sensitive to the possibility that the experimenter could infer choices (Hoffman et al., 1994; Hoffman, McCabe and Smith, 1996) and many studies have documented that what looks like *giving* can often be better classified as *giving-in* to social pressure (Cain, Dana and Newman, 2014).

In situations where people are directly confronted with pro-social requests, many follow reluctantly to avoid the feeling of shame. A reluctance to enter sharing environments has been observed in several field and laboratory studies. In a door-to-door fundraising study, DellaVigna, List and Malmendier (2012) find that informing households about an upcoming donation request significantly reduces the share of households opening the door. Dana, Cain and Dawes (2006) as well as Lazear, Malmendier and Weber (2012) document the same pattern in laboratory experiments where subjects are willing to (silently) sort-out of a dictator game at a cost.

Rather recently, scholars have started to investigate the robustness of several concepts which have previously been thought of as resulting from intrinsic preferences. Malmendier, te Velde and Weber (2014) find that a preference for reciprocating others' kindness is weaker than previously thought when social pressure is accounted for. Another example is presented by Kriss, Weber and Xiao (2016) who observe that third-parties punish norm violators reluctantly, i.e., although they indicate a preference for punishment, they ultimately avoid the act of punishing if excuses allow them to do so without blame. Attributing responsibility to nature allows subjects to maintain a positive image in the eyes of other subjects and the experimenter.

²There is an advanced literature connecting image motivation to lying and cheating (see e.g. Mazar, Amir and Ariely, 2008; Greenberg, Smeets and Zhurakhovska, 2015; Hao and Houser, 2017; Gneezy, Kajackaite and Sobel, 2018; Bicchieri, Dimant and Sonderegger, 2020).

One of the aims of our paper is to assess the role that social-image concerns play in promise keeping. We are only aware of few studies which have approached this territory before us. Deck, Servátka and Tucker (2013) hypothesise that the effectiveness of promises observed in earlier studies could partly be driven by subjects' image concerns towards the experimenter. The authors however cannot conclude this from their data because they could not replicate an effect of communication on cooperation under *both* single-blind and double-blind payoff procedures; a result which could be driven by ceiling effects as acknowledged by the authors. Schütte and Thoma (2014), on the other hand, test for social-image concerns by varying the ex-post observability of a promisor's action and report a null result. Again, a ceiling effect – this time stemming from the very high proportion of subjects keeping their promise in their baseline condition (81%) – could have limited the scope for image concerns to be detectable. Cadsby et al. (2015) find mixed evidence; they observe identifiability to matter for promise keeping in China, not however in New Zealand.

Our paper adds to the aforementioned literature by presenting the results of an experiment which is less susceptible to ceiling effects and which thereby provides a new testing ground for the relevance of social-image concerns in promise keeping.

1.2.3 Self-Image Concerns

Distinct from *social*-image concerns as outlined before are *self*-image concerns. People like to think of themselves as fair and honourable beings and where these perceptions are at stake, as in the case of opportunistic temptation, so is their *self-concept*. While psychologists have long recognised the importance of self-image concerns (e.g., Baumeister et al., 1998; Bem, 1972; Fiske, 2018), economists have only recently incorporated these concerns into what could be called “third-generation” theories of moral behaviour. Theories of self-concept maintenance (Mazar, Amir and Ariely, 2008), self-signalling (Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2006; Grossman and Van der Weele, 2017) and identity (Akerlof and Kranton, 2000, 2010; Bénabou and Tirole, 2011) have been able to organise findings unexplained by standard theories of social preference. Identity management, in particular *self-deception*, can explain why people avoid costless information (Dana, Weber and Kuang, 2007), sort-out of morally demanding situations (Dana, Cain and Dawes, 2006; Lazear, Malmendier and Weber, 2012), trade off good deeds with bad deeds (Mazar and Zhong, 2010; Merritt, Effron and Monin, 2010), or delegate the execution of opportunistic decisions to third-parties (Hamman, Loewenstein and Weber, 2010).

The importance of self-image concerns is also emphasised in a seminal study by Dana, Weber and Kuang (2007). In their “plausible deniability” treatment, a dictator can choose between an allocation favouring him- or herself over the recipient, or an equal and efficient allocation. The twist in this treatment is that the dictator can lose

agency if he or she delays making a decision in which case the computer intervenes to implement either outcome with equal chance. The recipient can never tell whether a selfish outcome resulted from a wilful decision or an unlucky dictator. Interestingly, delegating the choice to the computer is inconsistent with purely outcome-based theories of behaviour because such delegation would imply that the dictator prefers a lottery over two outcomes over each one separately. Self-image concerns, instead, become a natural candidate for explaining dictators’ willingness to delegate the decision. With 50% probability, the computer would choose the fair outcome the dictator would have felt compelled to choose anyway, but otherwise the selfish outcome would obtain and the dictator could maintain the illusion of not being responsible for its implementation. Indeed, a substantial proportion of dictators in their study (24%) allowed themselves to be cut off, thereby avoiding to make a decision.³ The deniability mechanism has further been applied to the analysis of reciprocal preferences e.g. by Van der Weele et al. (2014) and Regner (2018).

In our paper, we implement a variant of the cut-off mechanism to investigate the relevance of self-image concerns in promise keeping. To the best of our knowledge, all studies on promise keeping make it perfectly transparent to the decision maker that he or she is responsible for a broken promise; put differently, promise breaking is an act of commission. Yet, the responsibility for a broken promise can also be shifted to external circumstances, thereby granting a decision maker a moral excuse for selfish behaviour without compromising his or her self-image.

1.3 The Experiment

1.3.1 Design

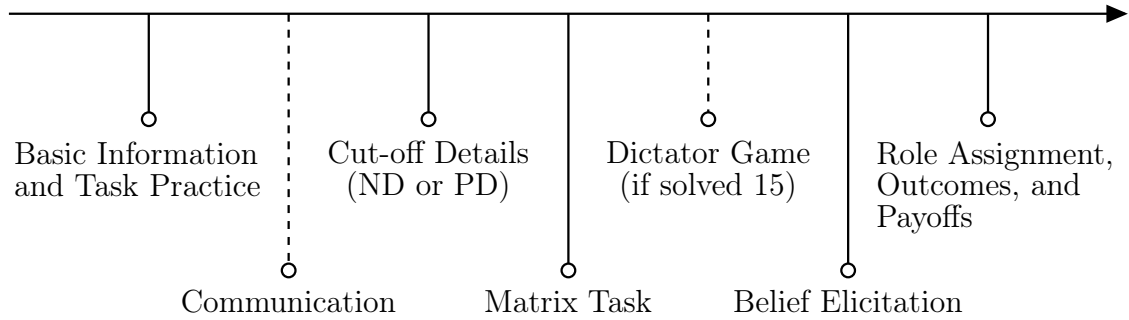
We combine a binary dictator game with a matrix solving task and systematically vary between subjects (i) the degree to which a “plausible deniability” mechanism allows subjects to obfuscate responsibility for outcomes and (ii) whether or not the experiment features a communication stage. Table 1.1 summarises our 2x2 factorial treatment design. The sequence of stages in the experiment is depicted in Figure 1.1.

Table 1.1: Factorial Treatment Design

	No Deniability	Plausible Deniability
No Communication	NC_ND	NC_PD
Communication	C_ND	C_PD

³Note that the cut-off timer was calibrated in a way such that subjects who really wanted to make a decision themselves had enough time to do so.

Figure 1.1: Sequence of Stages in the Experiment



Subjects are randomly paired in groups of two. Role assignment takes place *at the end* of the experiment, i.e., all subjects simultaneously play as *A* players (potential dictators) knowing that outcomes in this role would only count for half of them whereas the other half would eventually serve the role of player *B* (recipient).⁴

All treatments have in common that the dictator game stage is only reached if a preceding matrix task is solved successfully. In case of *success*, the subject enters the dictator game stage and decides how to allocate money between him- or herself and his or her counterpart by choosing one of two possible allocations: $A=(£10,£0)$ or $B=(£6,£6)$. Conversely, in case of *no success*, the subject skips the dictator game stage and is forced to let the computer randomly implement either of the two allocations with equal probability on his or her behalf.

The matrix task, borrowed from Abeler et al. (2011), consists of subjects counting *ones* (1s) in a series of 5x5 matrices comprised of randomly ordered zeros and ones.⁵ Importantly, we modified the task to feature a cut-off mechanism which (in some of our treatments) can serve as a plausible excuse for the implementation of the selfish allocation *A* (£10, £0).⁶ Successful completion requires a subject to solve a target amount of 15 matrices *on time*, i.e. before being cut off by the computer.

We employ different variants of the cut-off mechanism in our experiment. In our *No Deniability* (ND) treatments (Table 1.1, first column), subjects are given 300 seconds (5 minutes) to work on the task until a cut-off occurs. The time allotted in these treatments is extremely generous based on the results of an informal and unincentivised pretest where subjects needed on average 104s to solve 15 matrices and no subject took longer than 138s. Our aim was to erase the opportunity of using

⁴In the instructions, we refer to “you” and “your counterpart” instead of “dictator” and “recipient”. Instructions can be found in Appendix 1.B.

⁵Appendix 1.B.2 provides screenshots of the experimental interface.

⁶Recall that in Dana, Weber and Kuang (2007), 24% of the subjects allowed themselves to be cut-off by the computer, thereby preferring a mixture of two outcomes over each one separately. This observation is “inconsistent with a theory of rational choice with utilities defined only over outcomes” (p. 74). For subjects who are feeling compelled to choose the other-regarding option in order not to threaten their self-image, however, being cut off can be desirable. In half of the cases, the outcome would obtain which the dictator would have felt compelled to choose anyway. In another half of the cases, the opportunistic outcome would obtain allowing the subject to uphold the illusion of not being responsible for its implementation.

the cut-off mechanism as a plausible excuse for selfish allocations whilst keeping the experimental protocol as close as possible to the treatments we describe next.

In our *Plausible Deniability* (PD) treatments (Table 1.1, second column), instead of telling subjects that the cut-off would occur after 300 seconds sharp, we tell them that the cut-off can occur at any randomly determined second within the 300 seconds interval.⁷ The PD treatments offer room for two distinct dimensions of deniability:

- *Deniability towards the counterpart.* Subjects can exploit the fact that their counterpart cannot ascertain whether an outcome came about by a subject's own choice or by the computer. Our plausible deniability treatments therefore alleviate the social-image cost that is usually associated with selfish behaviour under full transparency.
- *Deniability towards the self.* Subjects who feel compelled to choose the generous allocation because they do not want to think badly of their selves may prefer to be cut off by the computer. A cut-off results in a fair chance (50%) of obtaining the opportunistic outcome whilst allowing to maintain the illusion of not being responsible for its implementation.

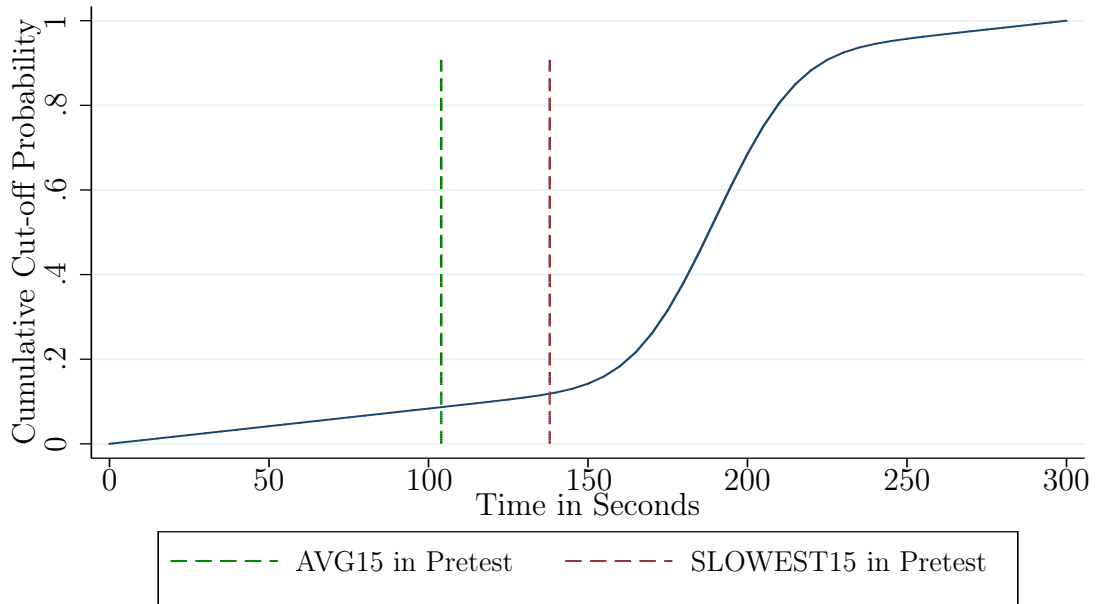
We assumed that self-deceivers would work on the task half-heartedly, waste time, or commit more errors all of which delaying the completion of the task.⁸ To identify whether subjects in our PD treatments indeed procrastinated, an additional control treatment was conducted. This treatment was designed as closely as possible to the NC_PD treatment. The only difference was the absence of a counterpart. In this treatment, successful completion of the matrix task allowed dictators to choose their own payoff only (£10 or £6). Since any incentives for procrastination were removed in this treatment, we aimed to obtain an unbiased distribution of performances in the matrix task against which to compare performances in our main treatments. Instructions for the control treatment can be found in Appendix 1.B.4.

No information was disclosed to subjects regarding the underlying distribution that generated the cut-offs in our PD treatments (and the control). Whilst it is technically true that a cut-off could occur anywhere within the specified time interval, we used a distribution which favoured later cut-offs. To be precise, we

⁷If a cut-off occurred, a subject was asked to work on a follow-up task for the remainder of the 300 seconds. The task was not incentivised and consisted of adding up numbers on screen. The purpose of this task was to maintain a constant sound of mouse clicks in the background, thereby ruling out that subjects could infer from the lack of this sound information about the timing of cut-offs of their peers.

⁸Previous studies which utilised a cut-off mechanism required self-deceivers to be passive and to wait for the computer to intervene. We decided to embed our cut-off mechanism into a real effort task instead of the dictator game itself to reduce potential demand effects and to mimic a richer (and in our opinion, more realistic) environment that would allow subjects to hide their intentions in an inconspicuous way, by disguising their true ability in an active task.

Figure 1.2: Calibrated Cut-off Distribution



combined a discretised normal distribution with a uniform distribution such that cut-offs would be drawn from the function: $f(x) = \mathcal{N}(190, 20) + \mathcal{U}\{1, 300\}$.⁹ Figure 1.2 depicts the associated cumulative distribution function which illustrates the probability of being cut off in the matrix task as a function of time. Dotted lines mark the times that the average as well as the slowest subject took to successfully complete the matrix task in the pretest. These times were used as benchmarks for our calibration. We calibrated the cut-off distribution with the following two objectives in mind:

- *Minimising data loss.*

Early cut-offs are associated with data loss because neither is the time data of a particular subject rich enough to identify procrastination nor do we obtain choice data in the subsequent dictator game. To minimise data loss, our cut-off distribution is shifted to the right. Recall that in the pretest, subjects needed on average 104s to succeed in the matrix task. But even up to the 150 seconds mark, the cumulative probability of being cut off in our experiment was merely 12% (after which it increased more rapidly).

- *Minimising selection effects.*

Some of the hypotheses derived in Section 1.3.2 are tested by comparing aggregate choice behaviour in the dictator game stage between our ND and PD treatments. For these tests to be reliable, we have to rule out the possibility

⁹We refrained from shifting the distribution to the utmost right and added a uniformly distributed element to it to preclude subjects from working out the underlying distribution ex-post e.g. through communication with fellow participants.

that our cut-off mechanism changed the composition of our PD compared to our ND samples. This would be the case e.g. if one assumed cut off subjects to be overly selfish or other-regarding. The shift of our cut-off distribution was specifically motivated to handle this potential concern. Since, for most of the cases, a cut-off would not occur until very late, we made it very difficult for subjects to successfully self-deceive. A cut-off could only be enforced through excessive procrastination which we assumed to be incompatible with maintaining the perception of irresponsibility. Consequently, we expected most subjects in our experiment to finish the task (with only few being cut off). In Section 1.4.2 we confirm that this was indeed the case in our experiment.

On the second dimension of our factorial treatment design, we varied whether subjects could communicate with their counterpart before entering the matrix solving stage. In the communication stage, we allowed subjects to exchange pre-formulated messages. Within a group, one subject was randomly chosen to send the first message by choosing one of the following alternatives:

Message 1: “I promise to do my best to implement Option B, if you promise to do the same.”

Message 2: “I don’t want to commit myself to anything.”

The second subject could then reply by choosing between:

Message 1: “I promise to do my best to implement Option B.”

Message 2: “I don’t want to commit myself to anything.”

Payoffs were calibrated providing an equality as well as total earnings maximising argument in favour of option B (£6, £6) over the opportunistic option A (£10, £0). We presumed that subjects would use the communication stage to exchange promises as a means to achieve cooperation on the former allocation.

The experiment was designed such that our deniability manipulations took place only after the communication stage had concluded. This means that, at the time when subjects exchanged messages, they did not know whether they would be assigned to the *No Deniability* or *Plausible Deniability* condition. It was only after messages had been exchanged and the communication stage had concluded that they learned which condition applied to them.¹⁰ By this means, we were able to vary by treatment whether deniability was possible or not without systematically influencing the content of exchanged messages.

¹⁰In the instructions, we only provide minimal information about the cut-off mechanism. Subjects are told that additional details would follow in the later course of the experiment. After the conclusion of the communication stage, treatment-specific details regarding the cut-off mechanism were read out aloud by the experimenter. Scripts can be found in Appendix 1.B.5.

By comparing the marginal effect of *adding* communication (and thereby promise exchange) to our existing ND and PD conditions, our experiment allows to shed light on the relevance of image concerns particular to promise keeping. We also collected data on subjects' beliefs about the behaviour and expectations of their counterpart to investigate whether any observed effects of our treatment variables are correctly anticipated by subjects to affect behaviour more generally. Subjects' second-order beliefs which serve as the conventional measure of guilt in the literature are moreover informative in assessing the role played by guilt aversion as a motivation for behaviour in our experiment.

Belief elicitation took place after the conclusion of the dictator game stage, but before roles and payoffs were assigned. Table 1.2 reproduces what subjects saw on their screen. Subjects were first asked how likely they thought it was that their counterpart (i) succeeded in the matrix task, and (ii) chose the generous allocation (conditional on having succeeded). Subsequently and on a separate screen, we elicited subjects' second-order beliefs by asking them to second-guess the responses of their counterpart to the aforementioned questions. Subjects were paid a flat payment of £1 for providing their initial responses. We decided not to incentivise the accuracy of these responses because the conventional approach would have required us to reveal information on a counterpart's true behaviour (which our PD conditions were specifically designed to avoid). This constraint did not apply to the elicitation of second-order beliefs which were formed upon a counterpart's beliefs rather than his or her actions. Consequently, we incentivised the accuracy of subjects' second-order beliefs by awarding a bonus of £1 for every response that was correctly matched.

Table 1.2: Belief Elicitation

How likely do you think it is that your counterpart correctly solved 15 matrices on time?					
	Very Likely	Somewhat Likely	50-50	Somewhat Unlikely	Very Unlikely
Your Guess	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Now, assume your counterpart correctly solved 15 matrices on time and made a choice between Options A and B. How likely do you think it is that your counterpart chose Option B (£6, £6)?					
	Very Likely	Somewhat Likely	50-50	Somewhat Unlikely	Very Unlikely
Your Guess	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

We opted for a one-shot version of the game because we presumed that learning associated with repeated play would eventually reduce or even erase the scope for self-deception to be operative. To assist subjects in their understanding of the rules and processes of the experiment, we initiated a practice phase in which they were guided through the stages of the experiment, supplemented with detailed on-screen explanations. In the course of this practice phase, subjects were also able to work on scaled-down versions of the matrix task with computer simulated counterparts. A late cut-off round (60s) familiarised them with how the matrix task worked, followed by an early cut-off round (12s) which was meant to familiarise subjects with the cut-off mechanism and its consequences.¹¹ The practice phase concluded with a quiz to ensure that subjects understood the instructions and processes of the experiment.

1.3.2 Hypotheses

We start this section by stating a set of more general hypotheses about the contents and effects of exchanged messages before turning our attention to *image motivation* in particular.

Hyp. 1: *Subjects will use the communication stage to exchange promises.*

Since the focus of our paper is on promise keeping, it was our intention to induce high rates of promise exchange in our experiment. Although some subjects may want to avoid commitment¹², we expected promise induced cooperation on the other-regarding allocation to be appealing to many subjects due to its equal and total-earnings maximising payoff property. Moreover, our restrictive communication protocol with pre-formulated messages made promise exchange suggestive and erased any ambiguities surrounding the classification of messages often observed under protocols of free form communication.

Hyp. 2: *Generosity is higher in treatments featuring communication.*

It is a well-documented finding in the literature that promises are often kept, even in one-shot encounters and in the absence of punishment threats. According to the *commitment-based explanation* of promise keeping, people keep their promises because they have an intrinsic preference for keeping their word. Consequently, we would expect some promise keeping to occur (and thereby increase generosity) under

¹¹To make it more apparent to subjects that a cut-off could be desirable, we programmed the computer to pick the opportunistic outcome in the early cut-off round. Thus, every subject experienced at least once that a cut-off could result in the implementation of the opportunistic outcome on the subject's behalf.

¹²Think of subjects who prefer keeping promises but expect their counterpart to make opportunistic promises which are bound to be broken. It is then rational for a subject not to engage in mutual promise exchange.

both our No Deniability *and* Plausible Deniability conditions.

Hyp. 3: *Beliefs about generosity are higher in treatments featuring communication.*

Hypothesis 3 naturally follows from hypothesis 2 under the assumption that subjects believe the underlying theory. It is the process by which promises feed expectations which also underlies the *expectations-based* explanation of promise keeping based on guilt and according to which people dislike letting others down on their promise-induced expectations.

We next turn our attention to understudied explanations of promise keeping which rest on the relevance of social- and self-image concerns. We contribute to the literature by assessing the empirical relevance of these explanations in our experiment.

1.3.2.1 Social-Image Concerns

From the stream of research discussed in Section 1.2.2, we know that subjects care about how they and their actions are being *perceived by others*. The assumption is that being perceived in a negative light by others imposes a psychological cost on the subject. Recall that the cut-off mechanism in our Plausible Deniability conditions could serve as an excuse for selfish outcomes. Since a subject's counterpart cannot ascertain how an outcome came about, we would expect social-image concerns to be mitigated in these treatments. Conversely, subjects in the No Deniability conditions cannot use early cut-offs as excuses for selfish outcomes. Therefore, we would expect social-image concerns to be amplified in these treatments.

The image concern that we are interested in arises over promise keeping. To rule out an alternative image concern, namely that of being perceived as *selfish* (or, greedy, unfair), we also conducted treatments where communication opportunities were removed. Our identification strategy is to compare the relative effectiveness of *adding* communication within our No Deniability as compared to our Plausible Deniability conditions.¹³ Under the assumption that there exist subjects who suffer an image cost of being perceived as a promise breaker by others, we would expect communication to be more effective under No Deniability compared to Plausible Deniability.

Hyp. 4: *Communication increases generosity more strongly under ND than PD.*

Again, given that subjects believe the underlying theory behind hypothesis 4, they will anticipate social image concerns to be amplified in others under ND compared to PD. We can state the following hypothesis:

¹³A similar strategy was applied by Schütte and Thoma (2014) in the context of a trust game.

Hyp. 5: *Communication increases beliefs about generosity more strongly under ND than PD.*

1.3.2.2 Self-Image Concerns

Our last set of hypotheses derive from the literature on self-image concerns which we discussed in Section 1.2.3. The message of this stream of research is that people desire to *perceive the self* in a favourable light. Psychological discomfort can be experienced when behaviour threatens a person's self-concept. One way of maintaining a desired self-concept in light of opportunistic temptation is to engage in self-deception.

Our idea is that self-image concerns may be relevant for promise-keeping. As a consequence, the strength of promises may be diluted in environments which allow people to self-deceive about the existence of a broken promise. In our experiment, a subject who feels compelled to live up to her promise in order not to threaten her self-image may want to procrastinate in the matrix task in the hope of being cut off by the computer. A cut-off results in a fair chance of obtaining the opportunistic outcome whilst allowing to maintain the perception of not having acted against one's promise. Recall that we conducted a control treatment where no counterpart was involved and successful completion of the matrix task allowed the dictator to choose her own payoff only. The assumption behind this treatment was that image related incentives for procrastination would be removed, thereby allowing us to obtain an unbiased approximation of subjects' ability in the matrix task against which to compare performances in our Plausible Deniability treatments (where we assumed such incentives to be present).

As argued before, image concerns can relate to *outcomes* (perceiving the self as selfish) and/or the *process* by which outcomes are reached (perceiving the self as a promise breaker). Considering our No Communication conditions first where only the former concern was at stake, we would expect self-deceivers in treatment NC_PD to have worked significantly more slowly and/or to have committed more mistakes compared to subjects in our control treatment.

Hyp. 6: *Matrix task performance is worse under NC_PD than CONTROL.*

In treatment C_PD, we assume that the additional self-image concern stemming from promise making induces higher generosity. This provides yet more subjects with an incentive to self-deceive and to procrastinate in the matrix task. From this, we predict matrix task performance in treatment C_PD to be worse compared to treatments NC_PD (and CONTROL).

Hyp. 7: *Matrix task performance is worse under C_PD than NC_PD.*

Recall that beliefs about generosity are expected to be higher in conditions featuring a communication stage. If anything, guilt aversion would therefore predict *more* instead of less effort in the matrix task which would bias our results *against* hypothesis 7.

1.3.3 Procedures

The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted in the *Laboratory for Economic and Decision Research* (LEDR) at the University of East Anglia. A total of 254 participants recruited from the local student population took part in the study. We ran 16 sessions in March 2018, each of which lasting between 35-45 minutes, depending on the treatment. We ran more PD sessions to compensate for the small data loss expected to occur by early cut-offs. The number of sessions per treatment were: 3 x NC_ND, 3 x C_ND, 4 x NC_PD, 4 x C_PD, 2 x CONTROL. 16 subjects took part in each session, except for one NC_PD session where only 14 subjects turned up. Average earnings were £10, with a minimum earning of £4 and a maximum earning of £16 (including a £3 participation fee).

Upon arrival, participants were randomly assigned to computer terminals by drawing their desk number. Each computer was located in a separate cubicle which inhibited visual interaction or communication. Anonymity amongst participants was secured because at no point during or after the experiment did any participant receive identifying information about his or her peers. We also took great care in the instructions emphasising that the experimenter would not be able to link the generated data to any participant as a person. Participants received a hard copy of the instructions and were asked to follow along as the experimenter read the instructions out aloud. Clarifications were provided on an individual basis. Participants were asked to answer a set of five control questions after the completion of the practice phase to ensure that they understood the instructions and processes of the experiment. Two further control questions were displayed after details regarding the cut-off mechanism were publicly announced by the experimenter. The experiment concluded with a brief questionnaire asking for socio-demographic characteristics. Privacy was ensured during the payment phase by asking participants to individually collect their final earnings from an assistant at the end of the experiment.

1.4 Results

Section 1.4.1 looks at the communication contents of our experiment. Section 1.4.2 analyses the effects of communication, focusing on social-image effects in Section 1.4.2.1 and self-image effects in Section 1.4.2.2.

1.4.1 Communication Contents

Table 1.3 summarises the observed message profiles (pairs of messages) broken down by treatment condition. Recall that by design, our deniability manipulations took place only after the communication stage concluded. Up to that point, the protocol of the experiment and the instructions were identical. We would therefore expect no significant differences in the contents of exchanged messages across treatments. This is confirmed by our data which is why we henceforth refer to the pooled data provided in the last column of Table 1.3.

By looking at the first two rows of Table 1.3, we can see that 46 out of 56 first-movers (82.1%) sent the cooperative message 1 stating a promise intent. Among the 46 second-movers who received a promise intent, 42 (91.3%) reciprocated with a promise thereby establishing mutual promise exchange. Unsurprisingly, amongst the few cases (10 out of 56) where first-movers refrained from proposing a mutual exchange of promises by stating that they do not want to commit themselves, the majority of second-movers (8 out of 10) decided not to commit either. Two subjects decided to commit despite not having received an intention to commit by their counterpart. In line with hypothesis 1, we can state the following result:

Result 1. *Most pairs of subjects (75%) used communication to exchange promises.*

Table 1.3: Overview of Message Profiles by Treatment

Message _{F-Mover} /Message _{S-Mover}	By Treatment			Pooled
	C_ND	C_PD	Z-stat. ^a (p-value)	C_ND + C_PD
Promise Intent/Promise	17/24 (70.8%)	25/32 (78.1%)	-0.624 (0.533)	42/56 (75%)
Promise Intent/No Commitment	3/24 (12.5%)	1/32 (3.1%)	1.348 (0.178)	4/56 (7.1%)
No Commitment/Promise	1/24 (4.2%)	1/32 (3.1%)	0.208 (0.835)	2/56 (3.6%)
No Commitment/No Commitment	3/24 (12.5%)	5/32 (15.6%)	-0.331 (0.741)	8/56 (14.3%)

^a The Z-statistic reflects two-tailed tests of differences in proportions.

1.4.2 Communication Effects

Having established that subjects used the communication stage to exchange promises, we can investigate *whether* and *by what means* promise exchange increased generosity in our communication treatments.

Our analysis is based on subjects who successfully completed the matrix task and for which choice data in the dictator game is available. Losing data on subjects who were cut off before the completion of the task may raise self-selection concerns. As discussed before, we designed our experiment to minimise these concerns. As expected, the proportions of subjects who were cut off in our Plausible Deniability conditions were small: 6/64 (9.4%) in treatment C_PD, 9/62 (14.5%) in treatment NC_PD, and 4/32 (12.5%) in treatment CONTROL. Moreover, if selection issues were present in the sense that procrastinators successfully managed to enforce a cut-off, we would expect the proportion of cut-offs to be higher in treatments C_PD and NC_PD (where incentives for procrastination were present) compared to treatment CONTROL (where incentives for procrastination were removed). This however was not the case according to pairwise Fisher's exact tests ($p = 0.441$ and $p = 0.529$ respectively, one-tailed). Appendix 1.A.1 provides details on cut-off times and matrix task progress of subjects who were cut off before they reached the target. It is noteworthy that a considerable proportion of these subjects (11/21 or 52.4%) did not manage to solve a single matrix in the practice stage, suggesting that our cut-off mechanism filtered out subjects who lacked a sufficient understanding of the task.

1.4.2.1 Social-Image Effects

All data referred to in this section is also subsumed in Table 1.4 which provides detailed summary statistics on the frequency of cut-offs, on choices in the dictator game stage, and on reported beliefs, all broken down by treatment and, if applicable, by communication history. Unless otherwise stated, reported Z statistics reflect tests of proportions (see Glasnapp and Poggio, 1985) when comparing choice data in the dictator game and Wilcoxon rank sum tests (see Siegel and Castellan, 1988) when comparing reported belief data.

Figure 1.3 summarises our main findings by depicting the proportions of subjects choosing the generous allocation for each treatment separately. Our communication protocol is effective in increasing generous allocations both under conditions of No Deniability (20.8% vs. 58.7%; $Z = -3.756$, $p < 0.01$, one-tailed) as well as under conditions of Plausible Deniability (18.9% vs. 37.9%; $Z = -2.215$, $p = 0.013$, one-tailed). A strong effect of communication is in line with hypothesis 2 and research discussed in Section 1.2.1. We state the following result:

Result 2. *Generosity is higher in treatments featuring communication.*

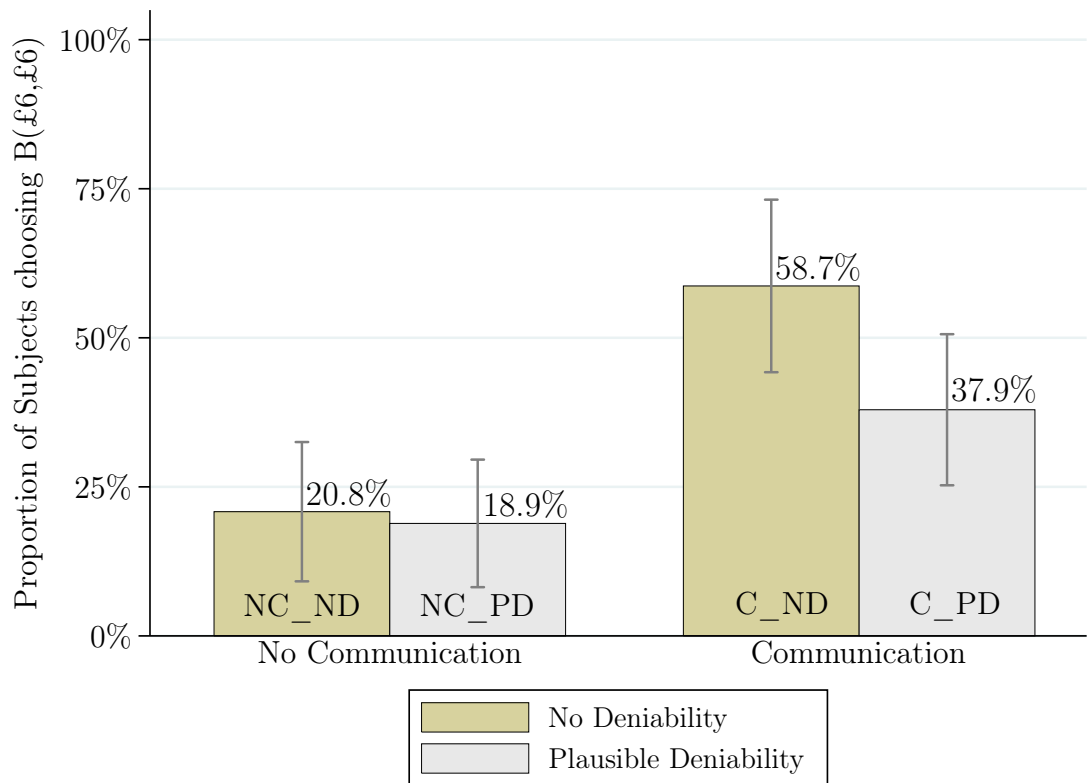
It is evident from our data however that communication has a stronger effect on generosity under ND compared to PD which squares well with hypothesis 4 and the idea that subjects dislike being perceived as a promise breaker by others.

Table 1.4: Summary Statistics

	n	Cut off n(%)	Generous n(%)	Selfish n(%)	Question 1 FO_Belief	Question 1 SO_Belief	Question 2 FO_Belief	Question 2 SO_Belief
Communication	112	8(7.1%)	49(47.1%)	55(52.9%)	4.44	4.46	2.88	2.88
C_ND	48	2(4.2%)	27(58.7%)	19(41.3%)	4.80	4.87	2.98	2.93
C_ND_PromiseEx.	34	1(2.9%)	25(75.8%)	8(24.2%)	4.79	4.85	3.24	3.21
C_ND_NoPromiseEx.	14	1(7.1%)	2(15.4%)	11(84.6%)	4.85	4.92	2.31	2.23
C_PD	64	6(9.4%)	22(37.9%)	36(62.1%)	4.16	4.14	2.81	2.84
C_PD_PromiseEx.	50	5(10.0%)	22(48.9%)	23(51.1%)	4.18	4.18	3.22	3.20
C_PD_NoPromiseEx.	14	1(7.1%)	0(0.0%)	13(100%)	4.08	4.00	1.38	1.62
No Communication	110	9(8.2%)	20(19.8%)	81(80.2%)	4.50	4.47	2.26	2.23
NC_ND	48	0(0.0%)	10(20.8%)	38(79.2%)	4.88	4.83	2.33	2.31
NC_PD	62	9(14.5%)	10(18.9%)	43(81.1%)	4.17	4.13	2.19	2.15
CONTROL	32	4(12.5%)	1(3.6%)	27(96.4%)	n/a	n/a	n/a	n/a

Note: Question 1 asked: “How likely do you think it is that your counterpart solved 15 matrices on time?”. Question 2 asked: “Now, assume that your counterpart solved 15 matrices on time and made a choice between Options A and B. How likely do you think it is that your counterpart chose Option B (£6, £6)?”. Answers were submitted on a 5 point Likert scale ranging from 1 (“very unlikely”) to 5 (“very likely”). A subject’s first-order belief (FO_Belief) is his or her response to these questions. A subject’s second-order belief (SO_Belief) is his or her guess regarding the response to these questions provided by their counterpart. Reported are the mean responses for each respective question.

Figure 1.3: Proportions of Generous Choices between Treatments



Result 3. *Communication increases generosity more strongly under ND than PD.*

As illustrated in Figure 1.3, our deniability manipulation affected generosity within our Communication conditions only, not however within conditions where no communication was possible. Looking at our Communication conditions first, we observe that plausible deniability significantly decreased the proportion of subjects choosing the generous allocation from 58.7% to 37.9% ($Z = 2.107$, $p = 0.018$, one-tailed). Considering promise keeping proportions in particular as shown in Table 1.4, we observe a significant decline from 75.8% in treatment C_ND to 48.9% in treatment C_PD ($Z = 2.396$, $p < 0.01$, one-tailed). Inspecting our No Communication conditions next reveals that plausible deniability decreased the proportion of generous allocations by merely two percentage points. Although the effect goes in the anticipated direction, the difference is insignificant ($Z = 0.248$, $p = 0.402$, one-tailed). It appears that subjects in the No Communication treatments were not particularly concerned about the transparency of their decisions. Or, put differently, purely outcome based image concerns (such as being perceived as selfish, egoistic, or unfair) seem not to have played a major role in our experiment. On the contrary, our results are compatible with the existence of a social image concern particular to promise keeping per se.

Do subjects predict the effects of our treatment variables on their counterpart's behaviour? We collected data on subjects' beliefs about their counterpart to answer this question. Responses were submitted on a 5 point Likert scale ranging from 1 (=“very unlikely”) to 5 (=“very likely”). As illustrated earlier in Table 1.2, we first asked subjects how likely they thought it was that their counterpart succeeded in the matrix task. The purpose of asking this first question was to check whether our deniability manipulations were successful in diffusing a counterpart's perceived responsibility for outcomes. As is evident from the data provided in Table 1.4, this was indeed the case. Plausible deniability decreased average first-order beliefs relating to question 1 within both our Communication (4.80 vs. 4.16; $Z = 4.623$, $p < 0.01$, one-tailed) and No Communication (4.88 vs. 4.17; $Z = 4.808$, $p < 0.01$, one-tailed) conditions. The same pattern holds for second-order beliefs. Allowing subjects to communicate, on the other hand, had no impact on a subject's belief about their counterpart's success in the matrix task.

Looking at first-order responses to question 2, we can see that communication and the exchange of promises raised subjects' *own* beliefs about a counterpart's generosity (2.26 vs. 2.88; $Z = -3.488$, $p < 0.01$, one-tailed). On top of that, communication was correctly predicted by subjects to also move their *counterparts'* beliefs about the subjects' own generosity as evidenced by subjects' second-order beliefs (2.23 vs. 2.88; $Z = -3.592$, $p < 0.01$, one-tailed). In line with hypothesis 3, we state the following result:

Result 4. *Beliefs about generosity are higher in our communication treatments. This suggests that subjects anticipated an effect of promise exchange on generosity.*

Comparing subjects' first-order responses to question 2 between our deniability conditions allows us to investigate whether subjects anticipated their counterpart to exploit the diffusion of responsibility inherent in our PD conditions. Relatedly, subjects' second-order responses are informative as to whether subjects anticipated their counterpart to anticipate such an effect to be present. In light of the fact that we did find an effect of deniability on behaviour as stated in result 3, it is surprising that subjects appear not to have anticipated deniability to matter to others. In the case of subjects' first-order beliefs (and equivalently so for second-order beliefs), we observe no statistical differences between our deniability conditions. This result holds both within our No Communication conditions (2.33 vs. 2.19; $Z = 0.762$, $p = 0.446$, two-tailed) and within our Communication conditions (2.98 vs. 2.81; $Z = 0.607$, $p = 0.544$, two-tailed).

Result 5. *The effect of communication on beliefs does not differ under ND and PD. This suggests that subjects failed to anticipate promise keeping to be sensitive to our deniability manipulations.*

It is interesting to see that guilt aversion – whilst providing a possible explanation (through result 4) for some of the generosity we observe – does not seem to capture the differences that we observe between our deniability manipulations. We observe higher generosity in treatment C_ND than C_PD despite there being no significant differences in subjects’ reported beliefs about generous behaviour between these treatments. Appendix 1.A.2 reproduces the results obtained in this section using regression analysis.

1.4.2.2 Self-Image Effects

Recall that despite being able to exploit deniability in treatment C_PD, a significant proportion of subjects (22/45 or 48.9%) honoured their promise. Conventional theories of promise keeping would argue that this effect is due to either an intrinsic preference for promise keeping (Vanberg, 2008), or an aversion to letting promisees down on their payoff expectations (Charness and Dufwenberg, 2006). Even social image concerns could still be present under the assumption that our PD treatments mitigated instead of fully erased perceived responsibility. An alternative explanation which has yet received little attention in the literature on promise keeping is that subjects honour their word to maintain their *self-image* as an honest person.

If self-image concerns contribute to the effectiveness of promises, we would expect its effect to be mitigated in environments which allow subjects to self-deceive about the cause of a broken promise. We hypothesised that self-deception in our experiment would take the form of subjects procrastinating in the matrix task to delegate their choice to the computer.

To obtain a benchmark for subjects’ abilities in the matrix task against which to compare performances in our plausible deniability treatments, we conducted our control treatment which erased incentives for procrastination. The following analysis is based on a comparison of performances in the matrix task observed between treatments C_PD, NC_PD, and CONTROL.

Table 1.5: Success Times and Accuracy in the Matrix Task

Treatment	n	Cut off n(%)	Time15 mean/median	Incorrect15 mean/median
NC_PD	62	9(14.5%)	102s/102s	1.49/1
C_PD	64	6(9.4%)	103s/100s	1.22/1
CONTROL	32	4(12.5%)	111s/104s	1.29/1

Figure 1.4: Cumulative Distribution Functions of Success Times

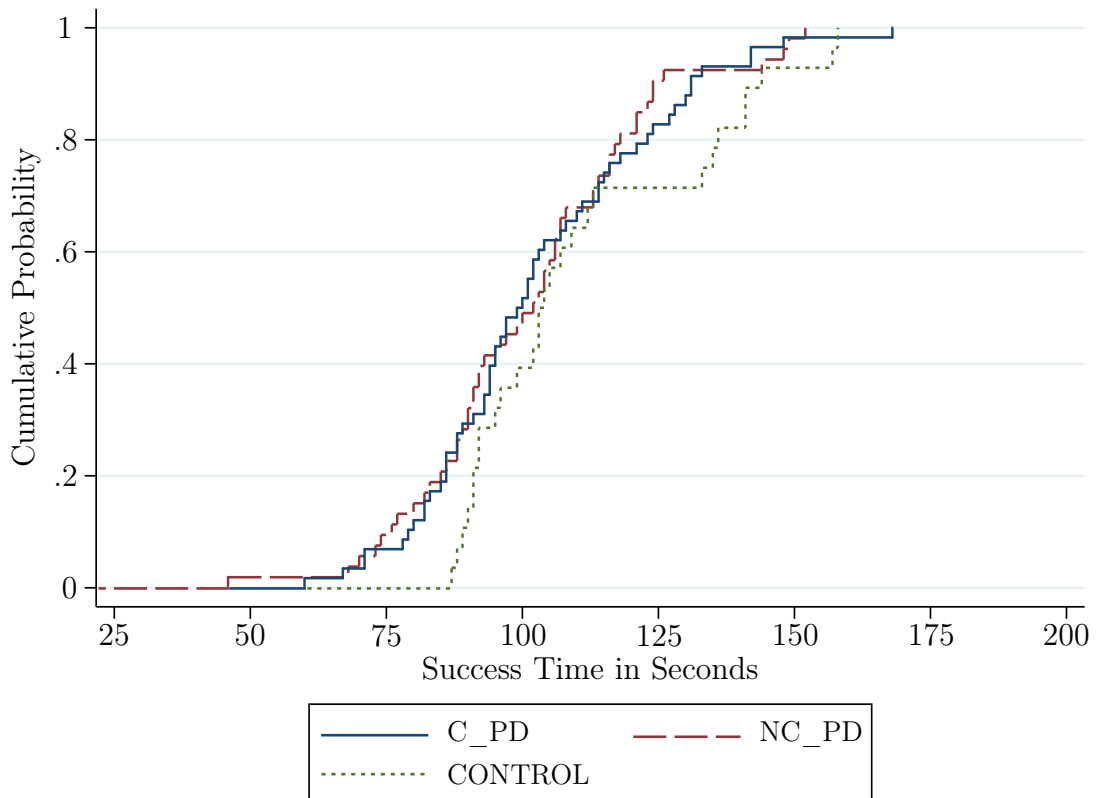
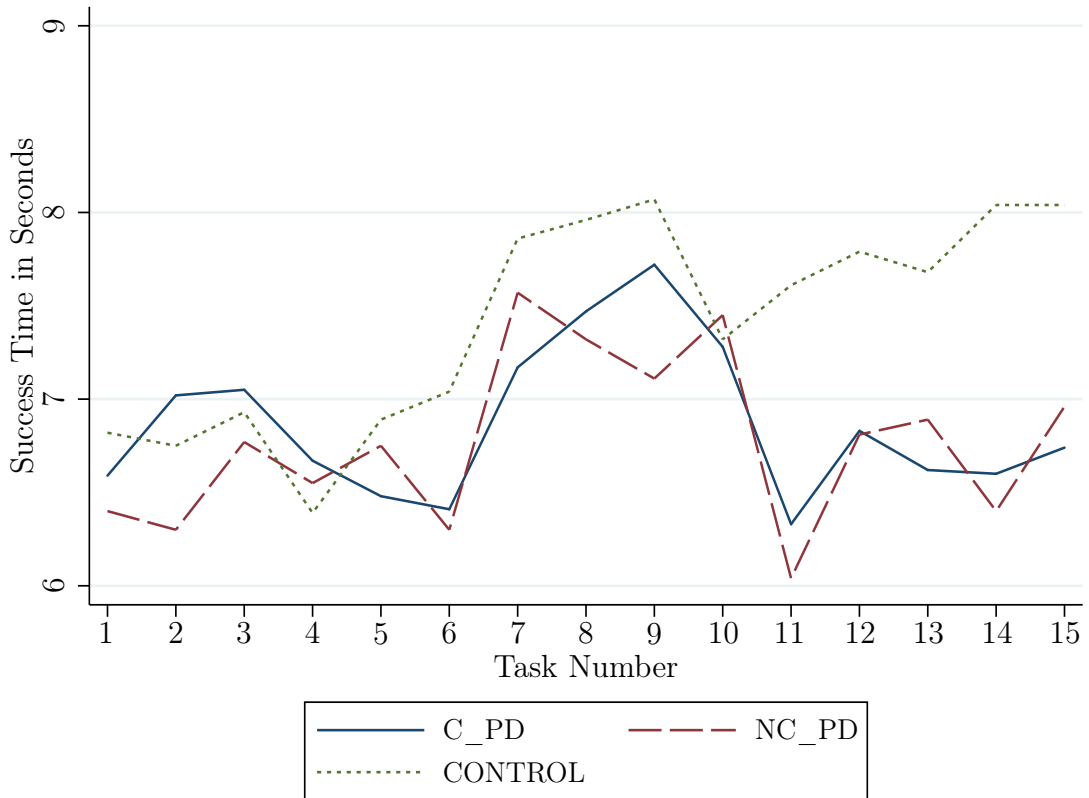


Table 1.5 reports summary statistics on the speed and accuracy with which subjects solved the target amount of 15 matrices.¹⁴ Figure 1.4 provides the associated cumulative distribution functions of success times across treatments. If subjects procrastinated in our main treatments, we would expect the respective CDF's to lie further to the right compared to our control treatment where incentives for procrastination were removed. We observe the opposite. It appears that subjects in our main treatments performed even better than subjects in the control which is particularly pronounced at the segment of high performing subjects. However, according to pairwise Kolmogorov-Smirnoff tests, the distributions of treatments C_PD and NC_PD do not differ significantly from CONTROL ($p = 0.157$ and 0.227 , respectively).

We also looked at within-subject variation of performances in the matrix task. It is possible that procrastination would take the form of subjects slowing down on the task the closer they approach the target amount of 15 matrices. Figure 1.5 depicts for every treatment separately the average time spent on each of the 15

¹⁴We continue to condition our analysis on the sample of subjects who have not been cut off. Recall our previous discussion on p. 17 and Appendix 1.A.1 for a justification of this approach. An advantage of doing so is that our cut-off mechanism simultaneously sorted out subjects who lacked sufficient understanding of the task. To keep these subjects in our sample would have made it complicated to discern motivated procrastination from delay due to misunderstanding.

Figure 1.5: Average Times Taken per Task



tasks. Again, eye-balling the results suggests that subjects in our main treatments performed better than subjects in the control treatment.

We ran a random effects panel model estimation to quantify what is observed in Figure 1.5. Results are presented in Table 1.6. Our dependent variable is the natural logarithm of the time (in seconds) taken by a subject to solve a given task. *TREAT* is a dummy distinguishing our treatment conditions with CONTROL being the reference category. *TASK_N* is the task number allowing us to measure changes in performance over time. We also include an interaction term between *TREAT* and *TASK_N* to allow performance changes to be treatment specific. The coefficient for *TASK_N* is positive and significant suggesting that subjects in our control treatment exhibit performance reductions as they move through the tasks. Such an effect could be due to boredom, or fatigue. On the contrary, no time trend is observed in treatments NC_PD and C_PD. This is evident from the negative coefficients of our interaction terms which are significant and fully compensate the negative time trend observed in our control treatment. Overall, performance in the matrix task appears to be worse in our control treatment with there being no difference between treatments C_PD and NC_PD. This result contradicts hypotheses 6 and 7 and lets us conclude with:

Result 6. *We find no evidence of procrastination in treatments NC_PD and C_PD.*

Table 1.6: Random Effects Panel Model Estimations

Dep. Variable <i>LN_TIME</i>	Coef.	Robust ^a Std. Error	Z	p-value
<i>TREAT</i>				
<i>NC_PD</i>	-0.015	0.050	-0.30	0.762
<i>C_PD</i>	0.023	0.050	0.47	0.641
<i>TASK_N</i>	0.012	0.004	3.15	0.002
<i>TREAT</i> × <i>TASK_N</i>				
<i>NC_PD</i>	-0.010	0.005	-2.23	0.026
<i>C_PD</i>	-0.012	0.004	-2.82	0.005
<i>_CONS</i>	1.847	0.041	45.31	0.000
Prob > chi2				0.013
R-Squared				0.015
Number of Groups				139
Number of Observations				2085

^a Standard errors are clustered on the subject level.

1.5 Discussion

Similar to previous studies looking at the role of non-binding verbal commitments on cooperative behaviour, we observe a positive relationship. It is noteworthy that the effects that we observe originate from a rather reserved protocol of pre-formulated message exchange which is commonly perceived to be less powerful than free-form communication (see e.g. Charness and Dufwenberg, 2010). We ascribe this result to the nature of our experimental protocol. Since we generate promise exchange in a dictator instead of a trust game framework, our environment is less susceptible to reciprocity effects which usually generate significant rates of trustworthiness in baseline conditions and thereby limit the scope for treatment effects to be detectable. The idea for this design feature goes back to Vanberg (2008)’s random dictatorship game. Our results suggest that this protocol may be of interest to researchers who prefer to resort to pre-formulated message exchange without making compromises on the effectiveness of promises, or those who are concerned about “ceiling effects” in trust game studies.

A separate examination of the effect of communication under No Deniability as compared to Plausible Deniability revealed that promise keeping was sensitive to

whether a promisee could undoubtedly blame the promisor for outcomes. Note that the observed effect cannot be attributed to an intrinsic preference which underlies the commitment-based explanation of promise keeping. This theory predicts promise keeping to be independent of image concerns. Our analysis was moreover able to rule out alternative explanations such as an image concern of being perceived as selfish, or an aversion to guilt. We also judge it unlikely that our results were driven by experimenter observability or demand because (i) the presence of the experimenter was not altered between treatments, and (ii) our treatment manipulations required only subtle changes to the experimental protocol. Instead, our results square well with the hypothesis that promise keeping is partly driven by subjects' aversion to being perceived as a promise breaker by their counterpart.

An interesting finding is the observation that subjects do not appear to have anticipated their counterpart to be sensitive to our deniability manipulations. This is surprising, given that subjects themselves did respond to the increased transparency embedded in our ND conditions by keeping their promises more often. It is possible that the emotion of shame, whilst being an important factor of a subject's own decision making process, is underestimated to play as important a role in others' behaviour. Under this premise, de-biasing subjects seems to be a promising avenue to foster the successful initiation of relationships based on trust.

Albeit to a lesser extent, promises remained to be effective even within our PD conditions. Both the commitment-based and the expectations-based explanations of promise keeping provide potential candidates for explaining this finding and our experiment was not designed to discern the empirical relevance of these theories from one another. Instead, we focused on a plausible alternative explanation of promise keeping which stems from the idea that subjects keep their promises in order not to threaten their self-image. This theory gave rise to the hypothesis that subjects would engage in self-deception – which would take the form of procrastination in the matrix task – to hide a reluctance to keep promises. We tested this hypothesis and report a null result.

One way of interpreting our null result is to take it as corroborating evidence of the strength of promises: subjects did not self-deceive because they truly desired to live up to their promise. At the same time, our result may call into question the generalisability of evidence supporting self-deception in dictator game studies to morally richer environments, as similarly pointed out by Van der Weele et al. (2014, p. 262). The authors implement a cut-off mechanism to investigate the robustness of reciprocal behaviour and likewise report a null result. There are caveats in order here, however.

Firstly, Regner (2018) reports a positive result observing that subjects do use the cut-off mechanism to avoid reciprocating others's kindness under different payoff calibrations of the trust game. He points out that the lack of a treatment effect in

Van der Weele et al. (2014) could be attributed to a ceiling effect stemming from the high proportion of selfish decisions (62.5%) observed in their baseline. Whilst a ceiling effect could have also been at work in our No Communication treatments, it is less likely that the same applied to our Communication treatments where the proportion of selfish allocations in our ND baseline was merely 41.3%.

Lastly, it is important to point out differences in the way we designed our experiment as compared to the aforementioned studies and in particular compared to the seminal paper by Dana, Weber and Kuang (2007). Whereas in their study, self-deception required subjects to deliberately wait for the computer to intervene, in our study subjects could delegate their decision in a more subtle and inconspicuous way by means of procrastination in an *active* task. One could argue that our design is less susceptible to demand effects and therefore provides a more natural testing ground for self-deception. At the same time, our experiment is more complex. It is possible that the additional complexity of our experiment made it more difficult for subjects to fully process the “exploitability” of our cut-off mechanism. However, as discussed in the design section of our experiment, we initiated a practice phase to assist subjects’ general understanding of our game. In the course of this practice phase, we also exposed subjects to outcomes which hinted at the possible desirability of being cut off in our experiment.

1.6 Conclusion

Trust is often referred to as the glue to social capital formation. Although its efficiency enhancing nature is desirable, trust can also be betrayed. Communication and the exchange of promises are among the most prominent mechanisms to promote trust.

The experiment that we presented was specifically set out to assess the relevance of two understudied explanations of promise keeping, namely social- and self-image concerns. We observe evidence of social-image concerns in treatments which feature ex-ante opportunities for promise exchange. Ruling out alternative explanations, our results are consistent with subject exhibiting an aversion to being perceived as a promise breaker by others. Surprisingly, subjects seem not to anticipate social-image concerns to be present in others. Our test of self-image concerns yielded a null result: there is no evidence of subjects engaging in self-deception to evade their promise-induced commitments. This resilience could be interpreted as corroborating evidence of the strength of promises.

Our study contributes to the literature on promise keeping by documenting the significance of social-image concerns and, to the best of our knowledge, by being the first to have tested for self-image concerns.

Appendix for Chapter 1

1.A Supplementary Data

1.A.1 Cut-offs and Task Performance

Table A.1 lists subjects who were cut-off from the task before successfully solving the required number of 15 matrices. Overall, this was the case for 21 out of 254 (8.3%) subjects in our experiment. In the last column, we indicate whether or not a respective subject correctly solved any of the matrices of the practice phase of our experiment. This information may be informative as to whether or not a subject struggled understanding the task. For some subjects, this appears to have indeed been the case as is evident e.g. from subject #249 who made 96 mistakes in the control treatment. Another subject directly expressed to the experimenter confusion about how to solve a given matrix in the practice stage.

It appears that many of the cut-offs observed are consistent with delays due to misunderstanding rather than procrastination. Examining the reported cut-off times and the progress of subjects who demonstrated understanding of the task, it does not appear to be the case that cut off subjects were reluctant to solve the task.

Table A.1: Cut-offs and Task Performance

	Treatment	ID	Session	Cut-off Time	#Correct	#Incorrect	Solved Practice?
1.	NC_PD	52	9	176	12	4	No.
2.	NC_PD	57	9	23	3	0	Yes.
3.	NC_PD	60	9	193	0	6	No.
4.	NC_PD	61	9	25	5	0	Yes.
5.	NC_PD	72	10	35	3	1	No.
6.	NC_PD	73	10	65	9	0	Yes.
7.	NC_PD	101	12	38	1	4	No.
8.	NC_PD	103	12	113	12	2	Yes.
9.	NC_PD	104	12	22	3	0	Yes.
10.	C_ND	115	5	300	0	14	No.
11.	C_ND	152	7	300	0	7	No.
12.	C_PD	179	2	83	9	1	Yes.
13.	C_PD	182	2	78	10	0	Yes.
14.	C_PD	191	3	86	12	0	Yes.
15.	C_PD	195	3	41	5	0	Yes.
16.	C_PD	206	3	100	12	3	No.
17.	C_PD	220	4	165	6	8	No.
18.	CONTROL	225	8	175	13	3	No.
19.	CONTROL	242	16	66	4	2	No.
20.	CONTROL	246	16	106	14	1	Yes.
21.	CONTROL	249	16	171	2	96	No.

1.A.2 Regression Results

In Table A.2 we report supplementary regression results supporting the conclusions derived from our non-parametric analyses reported in the main text. The dependent variable in models [1]-[2] is a dummy taking value 1 if the generous allocation was chosen, and 0 otherwise. In models [3]-[6], the dependent variable is the respective question 2 belief measured on a 5 point Likert scale. As independent variables we include dummies for our treatment conditions and the interaction thereof. What we find is that communication exerts a strong influence on generosity *and* on reported beliefs which is consistent with the idea that an effect of communication could partly be mediated through guilt aversion. The negative interaction term in models [1]-[2] moreover suggests that communication exerts a stronger effect on generosity within our No Deniability conditions. Interestingly and in line with our previous findings, this asymmetry is not mirrored by beliefs which is evident from the insignificant interaction term reported in models [3]-[6].

The result that communication and promise exchange affect behaviour more strongly under No Deniability, coupled with the finding that beliefs were not affected, suggests that the effect of our deniability manipulations is unlikely to be attributed to an aversion to guilt. Instead, our findings are consistent with subjects exhibiting an aversion to being perceived as a promise breaker by others.

Table A.2: Regression Results

	[1]	[2]	[3]	[4]	[5]	[6]
Model:	Probit	OLS	Ord. Logit	OLS	Ord. Logit	OLS
Dep. Variable:	Generous	Generous	FO-Belief	FO-Belief	SO-Belief	SO-Belief
Communication	1.032*** (0.173)	0.379*** (0.053)	0.904** (0.408)	0.645** (0.272)	0.870*** (0.203)	0.622*** (0.155)
Pl. Deniability	-0.071 (0.202)	-0.020 (0.057)	-0.224 (0.445)	-0.145 (0.287)	-0.252 (0.355)	-0.162 (0.239)
Communication × Pl. Deniability	-0.456* (0.236)	-0.188** (0.074)	0.006 (0.542)	-0.023 (0.351)	0.121 (0.437)	0.072 (0.294)
Constant	-0.812*** (0.163)	0.208*** (0.047)		2.333*** (0.188)		2.313*** (0.141)
(Pseudo) R-Squared	0.084	0.108	0.022	0.074	0.022	0.075
n	205	205	205	205	205	205

Note: All regressions cluster observations on the session level. Robust standard errors are reported in parentheses. *(**), ***): coefficient significantly different from zero at the 10% (5%, 1%) level.

1.B Instructions and Screens

1.B.1 Main Treatment Instructions

Information in brackets [...] only applies to treatments featuring a communication stage. Otherwise, instructions are identical across our four main treatments. Subjects received information regarding the cut-off mechanism just before entering the matrix solving stage.

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

At the beginning of the experiment, you will be paired with another randomly determined participant in the room who will from now on be called your **counterpart**. No participant will get to know the identity of his/her counterpart during or after the experiment.

All participants in this experiment are provided with the same set of instructions and will encounter the same stages as described below:

Stage 1: Matrix Task.

In stage 1 of the experiment, you will work on a matrix solving task. The task consists of counting *ones* (1s) in a series of matrices comprised of random 0s and 1s. A sample matrix is depicted in Figure 1 below.

Figure 1: Sample Matrix

0	1	1	1	0
0	0	0	0	0
1	0	0	1	1
1	1	1	1	0
1	1	0	1	0

You will be able to work on this task for a *maximum* of 300 seconds (5 minutes). Importantly, you will be timed-out by the computer at some point during this time interval. If this happens, the matrix task will end. You will then be asked to work on a follow-up task for the remainder of the 300 seconds.

All participants will be provided with additional details regarding the time-out mechanism in the later course of the experiment.

Outcomes in the matrix task (not however in the follow-up task) have direct consequences for the decision environment in stage 2 of the experiment:

- If you correctly solve at least 15 matrices before you are timed-out by the computer, you will be able make a decision in stage 2 of the experiment.
- If you do not correctly solve at least 15 matrices before you are timed-out by the computer, you will *not* be able to make a decision in stage 2 of the experiment.

After the conclusion of the matrix and follow-up task (i.e. after 300 seconds), you will move forward to stage 2 of the experiment.

Stage 2: Decision Stage.

In stage 2 of the experiment, you will *potentially* be able to choose between two options. Your choice indicates how you would like to allocate money between you and your counterpart. The possible options are:

- Option A: **£10** to you and **£0** to your counterpart.
- Option B: **£6** to you and **£6** to your counterpart.

If you succeeded in solving at least 15 matrices in stage 1 of the experiment, *you yourself* will choose between Option A and Option B.

If you did not succeed in solving at least 15 matrices in stage 1 of the experiment, *the computer instead* will randomly choose between Option A and Option B with equal probability.

The resulting option (A or B) will be called your **individual stage 2 outcome**. You will know whether the outcome of your stage 2 was determined by your own choice or by the choice of the computer. Your counterpart, however, *will not* know how your stage 2 outcome came about.

Determining the Relevant Player.

After both you and your counterpart have individually completed the stages above, one of you will be randomly determined by the computer to become the **Relevant Player**.

If you become the Relevant Player, your stage 2 outcome will be implemented. If you *do not* become the Relevant Player, your stage 2 outcome *will not* be implemented and will therefore have *no consequences* for payoffs in the experiment. In this case, your payoffs will solely be determined by the stage 2 outcome of your counterpart because he or she was assigned the role of Relevant Player.

Note that it is *equally likely* that you or your counterpart will be assigned the role of Relevant Player.

[Communication Phase.

Before stage 1 of the experiment starts you will be asked to choose one of two pre-defined messages to be sent to your counterpart.

Note that at this point, you will not know which of you will become the Relevant Player in the experiment. You will receive this information only at the end of the experiment.

Messages will be exchanged sequentially. One participant will be randomly determined to send the first message by choosing one of the following options:

Message 1: *“I promise to do my best to implement Option B, if you promise to do the same.”*

Message 2: *“I don’t want to commit myself to anything.”*

The second participant in a group will then be asked to reply by choosing one of the following options:

Message 1: *“I promise to do my best to implement Option B.”*

Message 2: *“I don’t want to commit myself to anything.”*

Importantly, the sequence in which messages are exchanged is randomly determined and not related to the assignment of roles at the end of the experiment. *Again, this means that at the time when you exchange messages with your counterpart, you will not know which of you will be assigned the role of Relevant Player.]*

Bonus: Guessing.

At certain points during the experiment, you will have the opportunity to earn small amounts of additional money by guessing decisions and outcomes in the experiment. You will learn more about this during the experiment.

Practice.

We will now briefly guide you through the decision stages in order for you to get a better understanding of the interface and processes of this experiment. You will also be able to familiarise yourself with the matrix task. We will conclude the practice phase with a quiz to check your understanding.

Please follow along on screen.

1.B.2 Practice Stage Screens

Intro Screen

Practice
In this practice phase we will guide you through the decision stages of this experiment for you to get a better understanding of the interface and processes of this experiment. The box below will contain the screens that you will encounter in the experiment.
Every screen in this practice phase will also contain a white header box which may contain additional clarifications about the rules of the experiment.
You will be asked to confirm that you understood the information on top by clicking the 'Understood' button. Only then will the main screen below unlock.
Note that in this demonstration, your counterpart is simulated by the computer. You will be matched with a real participant once the main experiment starts.

Understood

Press the button below to start the practice phase.

Ready

Communication Screen

Practice
Communication Phase:
Before stage 1 of the experiment starts, you will be able to exchange messages with your matched counterpart.
If you were selected to send the first message, the box below shows how your screen will look.
You can choose one of the example messages from the box on the left-hand side and transmit it to you counterpart by confirming your message using the 'send' button.
Your message will then appear in the chat box on the right-hand side. Give it a try!

Understood

You are the first mover
Please select the first message to be sent

Message Box	Chat Box
<p><input type="radio"/> Hello there!</p> <p><input type="radio"/> Good luck to the two of us!</p> <p style="text-align: right;">Send</p>	<p>Me:</p> <p>My counterpart:</p>

Proceed

Communication Screen (cont.)

Practice
Communication Phase:
 If you were selected to send the second message instead, the box below shows how your screen will look.
 You can reply by selecting one of the messages from the box on the left-hand side and transmit it to your counterpart by confirming your message using the 'send' button.
 Your message will then appear in the chat box on the right-hand side. Give it a try!

Understood

You are the second mover
Please select a message to reply to your counterpart

Message Box

Hi, I hope you are well!
 Good luck!

Send

Chat Box

My counterpart:
Good luck to the two of us!

Me:

Proceed

Matrix Task Intro Screen – Round 1

Practice Round 1
 You will now be able to work on a scaled down version of the matrix task.
 In order to be able to make a choice in stage 2, you have to solve at least 3 matrices within a maximum of 60 seconds (1 minute).
 You will encounter this task twice in this practice phase. The computerized time-out will differ in each of these two rounds.
 In the first encounter of this practice round, the computer will time you out after the maximum allotted time of 60 seconds (1 minute) has passed.
 We have deliberately chosen the time-out to occur late. This allows you to familiarize yourself with the matrix task. Give it a try!

Understood

The button below will start the matrix task.

If you correctly solve at least 3 matrices before you are interrupted by the computer, you will be able to choose between:

Option A: £10 to you and £0 to your counterpart
 Option B: £6 to you and £6 to your counterpart

If you **do not** correctly solve 3 matrices on time, the computer instead will randomly implement either Option A or Option B with equal probability.

Start the matrix task

Matrix Task Screen – Round 1 (Late Cut-off)

Practice Round 1

0	0	1	0	1
1	1	1	0	1
0	0	0	1	0
1	0	0	0	1
0	1	0	1	0

Correct: 3
Incorrect: 0

Answer

OK

Success Screen – Round 1

Practice Round 1

You have solved enough matrices on time!
Please select your preferred option.

- Option A:** £10 to you and £0 to your counterpart
- Option B:** £6 to you and £6 to your counterpart

Continue

Relevant Player Screen – Round 1

Practice Round 1

Your individual stage 2 outcome is:
Option B: £6 to you and £6 to your counterpart

The stage 2 outcome of your counterpart may be different from yours.
By clicking the button below, either you or your counterpart will be assigned the role of **Relevant Player**.
The likelihood of becoming the Relevant Player is the same for you and your counterpart.
Only the stage 2 outcome of the Relevant Player will influence earnings in the experiment.

Continue

Outcome Screen – Round 1

Practice Round 1

You have been assigned the role of Relevant Player!
This means your stage 2 outcome was implemented:

Option B: £6 to you and £6 to your counterpart

Your earnings:

Participation fee:	£3
Stage 2 earnings:	£6
Final earnings:	£9

Continue

Matrix Task Intro Screen – Round 2

Practice Round 2
 You will now be able to work on the task for a second time.
 In order to be able to make a choice in stage 2, you have to solve at least 3 matrices within a maximum of 60 seconds (1 minute).
 In this encounter, we have deliberately chosen the time-out to occur very early.
 This allows you to familiarize yourself with the time-out mechanism and its consequences.

Understood

The button below will start the matrix task.

If you correctly solve at least 3 matrices before you are interrupted by the computer, you will be able to choose between:

Option A: £10 to you and £0 to your counterpart
 Option B: £6 to you and £6 to your counterpart

If you **do not** correctly solve 3 matrices on time, the computer instead will randomly implement either Option A or Option B with equal probability.

Start the matrix task

Matrix Task Screen – Round 2 (Early Cut-off)

Practice Round 2

1	0	0	1	0
1	0	1	0	0
0	0	1	1	0
0	1	1	0	1
1	0	0	0	1

Correct: 0

Incorrect: 1

Answer

OK

Time-out Screen – Round 2

Practice Round 2
If you are timed-out, you will see the screen below.
The experiment will automatically continue once the maximum allotted time for stage 2 (1 minute) has elapsed.
In the meantime, we would like you to work on a follow-up task.
This task has no consequences for the outcomes or payoffs in the experiment.

Understood

You have been timed-out!

The experiment will continue soon.

In the meantime, we are interested in your performance in a follow-up task.
This task has no consequences for the outcomes or payoffs in the experiment.
You can start the task by clicking on the button below.

Start follow-up task

Filler Task Screen – Round 2

Practice Round 2

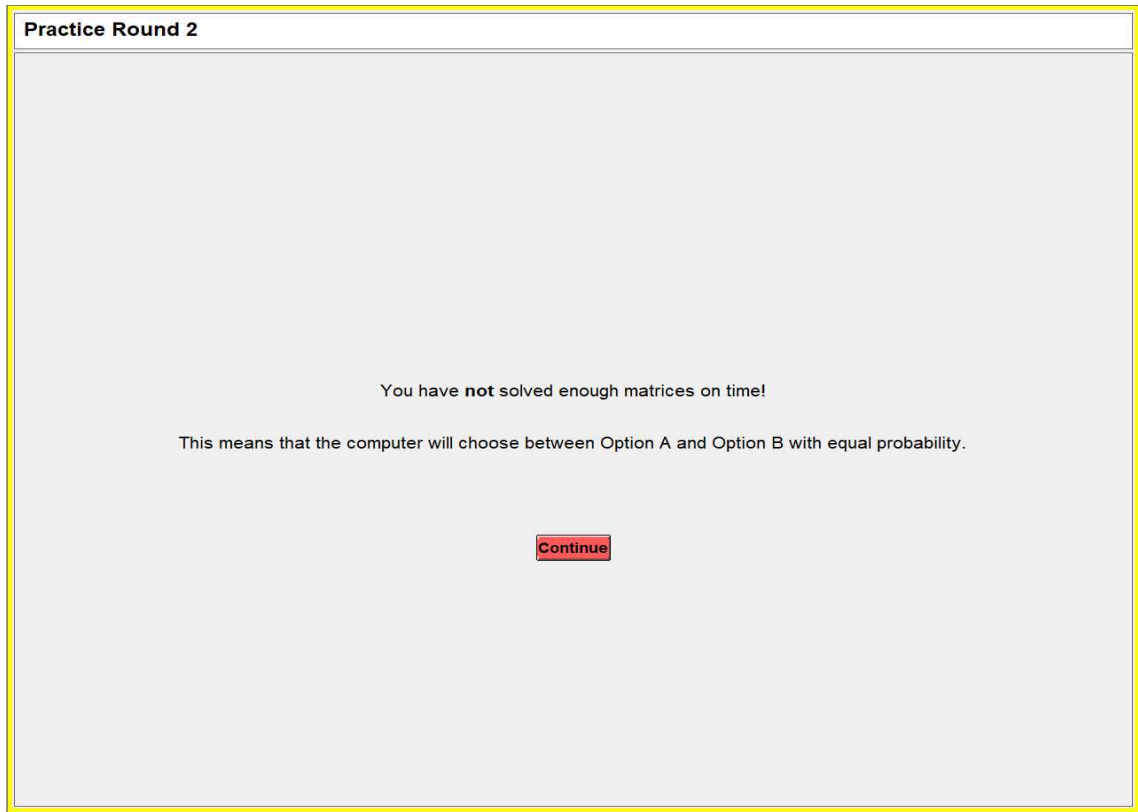
Correct:	0
Incorrect:	0

$40 + 83 + 1$

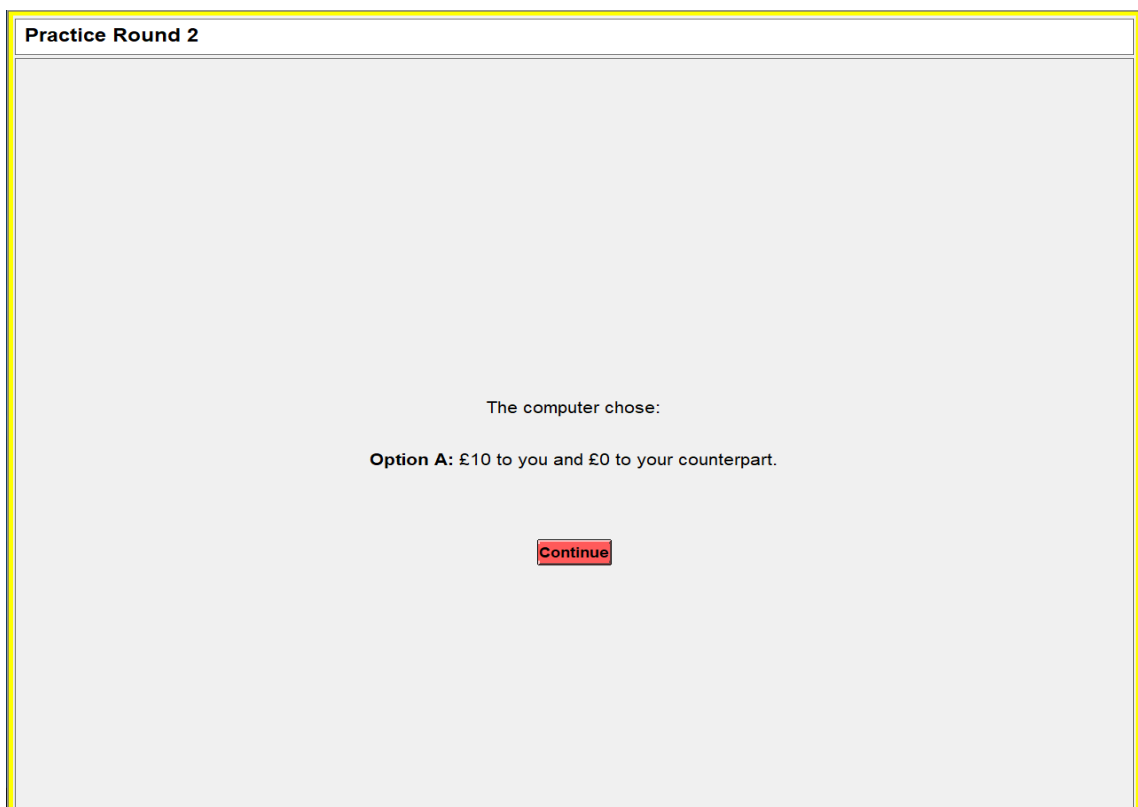
Answer

OK

No Success Screen – Round 2



Computer Randomisation Screen – Round 2



Relevant Player Screen – Round 2

Practice Round 2

Your individual stage 2 outcome is:
Option A: £10 to you and £0 to your counterpart.

The stage 2 outcome of your counterpart may be different from yours.
By clicking the button below, either you or your counterpart will be assigned the role of **Relevant Player**.
The likelihood of becoming the Relevant Player is the same for you and your counterpart.
Only the stage 2 outcome of the Relevant Player will influence earnings in the experiment.

Continue

Outcome Screen – Round 2

Practice Round 2

You have **not** been assigned the role of Relevant Player!
This means your counterpart's stage 2 outcome was implemented:

Option A: £0 to you and £10 to your counterpart

Your earnings:

Participation fee:	£3
Stage 2 earnings:	£0
Final earnings:	£3

Continue

1.B.3 Control Questions

*Upon completion of the practice stage, subjects were asked to answer a set of 5 control questions. 2 more control questions were assessed after details regarding the cut-off mechanism were announced. Questions in the control treatment differed only marginally which is why they are not explicitly reported here. Correct answers are highlighted. Where correct answers differ, **green** marks the correct answer in the No Deniability conditions and **blue** marks the correct answer in the Plausible Deniability conditions.*

Control Question 1:

The data generated in this experiment ...

- ✓ is anonymous, neither the experimenter nor other participants will be able to link my behaviour to me as a person.
- links my behaviour in the experiment to me as a person.
- links my behaviour in the experiment to me as a person, but only the experimenter will be able to make this connection.

Control Question 2:

Participants in this experiment ...

- are provided with different instructions but will encounter the same stages in the experiment.
- ✓ are all provided with the same instructions and will encounter the same stages in the experiment.
- will encounter different stages in the experiment.

Control Question 3:

I will be able to choose between Option A and Option B ...

- no matter what.
- ✓ only if I solve enough matrices on time.
- only if I will be timed-out by the computer in the matrix task.

Control Question 4:

My stage 2 outcome will contribute to my earnings in the experiment ...

no matter what.

- ✓ only if I become assigned the role of Relevant Player at the end of the experiment.

only if I become assigned the role of Relevant Player at the beginning of the experiment.

Control Question 5:

My counterpart ...

will learn whether I succeeded in the matrix task.

will learn whether my stage 2 outcome was chosen by me or by the computer.

- ✓ will neither learn my performance in the matrix task nor whether my stage 2 outcome was chosen by me or by the computer.

Control Question 6:

In this experiment ...

- ✓ I will be timed out when the maximum allotted time of 300 seconds is reached.

- ✓ I can be timed out at any second within the maximum allotted time of 300 seconds.

I will never be timed out.

Control Question 7:

What will your counterpart know after you completed the matrix task?

My counterpart will know how many matrices I solved.

- ✓ My counterpart will know that I was able to work on the matrix task for 300 seconds.

- ✓ My counterpart will know that I was timed out anywhere within 300 seconds. He will however not know when exactly my time out occurred.

1.B.4 Control Treatment Instructions

In this treatment, we erased the recipient role. All other features of this treatment including the instructions and experimental procedures closely followed treatment NC_PD. We also implemented a counterpart to the role uncertainty feature in the main treatments which meant that outcomes would only count in half of the cases. In the control treatment, we let the computer pick a ‘relevant scenario’ instead of a ‘relevant player’. If a subject’s scenario was determined not to count, a compensation of £3 was awarded which lies just in-between the two possible payoff allocations (£6 or £0) which a subject could have expected to be allocated in the main treatments by her counterpart.

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

All participants in this experiment are provided with the same set of instructions and will encounter the same stages as described below:

Stage 1: Matrix Task.

In stage 1 of the experiment, you will work on a matrix solving task. The task consists of counting *ones* (1s) in a series of matrices comprised of random 0s and 1s. A sample matrix is depicted in Figure 1 below.

Figure 1: Sample Matrix

0	1	1	1	0
0	0	0	0	0
1	0	0	1	1
1	1	1	1	0
1	1	0	1	0

You will be able to work on this task for a *maximum* of 300 seconds (5 minutes). Importantly, you will be timed-out by the computer at some point during this time interval. If this happens, the matrix task will end. You will then be asked to work on a follow-up task for the remainder of the 300 seconds.

All participants will be provided with additional details regarding the time-out mechanism in the later course of the experiment.

Outcomes in the matrix task (not however in the follow-up task) have direct consequences for the decision environment in stage 2 of the experiment:

- If you correctly solve at least 15 matrices before you are timed-out by the computer, you will be able make a decision in stage 2 of the experiment.
- If you do not correctly solve at least 15 matrices before you are timed-out by the computer, you will *not* be able to make a decision in stage 2 of the experiment.

After the conclusion of the matrix and follow-up task (i.e. after 300 seconds), you will move forward to stage 2 of the experiment.

Stage 2: Decision Stage.

In stage 2 of the experiment, you will *potentially* be able to choose between two options. Your choice indicates how much money you would like to allocate to yourself. The possible options are:

- Option A: **£10** to you.
- Option B: **£6** to you.

If you succeeded in solving at least 15 matrices in stage 1 of the experiment, *you yourself* will choose between Option A and Option B.

If you did not succeed in solving at least 15 matrices in stage 1 of the experiment, *the computer instead* will randomly choose between Option A and Option B with equal probability.

The resulting option (A or B) will be called your individual stage 2 outcome.

Determining the Relevant Scenario.

After you have completed the stages above, the computer will randomly determine whether your stage 2 outcome becomes the **Relevant Scenario**.

If your stage 2 outcome becomes the Relevant Scenario, it will be implemented. If your stage 2 outcome *does not* become the Relevant Scenario, your stage 2 outcome *will not* be implemented and will therefore have *no consequences* for payoffs in the experiment. In this case, you will instead earn a compensation of £3.

Note that it is *equally likely* that your stage 2 outcome *will* or *will not* become the Relevant Scenario.

Practice.

We will now briefly guide you through the decision stages in order for you to get a better understanding of the interface and processes of this experiment. You will also be able to familiarise yourself with the matrix task. We will conclude the practice phase with a quiz to check your understanding.

Please follow along on screen.

1.B.5 Revelation of Cut-off Details

Just before subjects entered the matrix solving stage, we publicly announced treatment specific details regarding the cut-off mechanism both verbally and on screen.

Script [1] for treatments NC_ND and C_ND:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a fixed point in time. You will be timed out when the maximum allotted time of 300 seconds (5 minutes) is reached. Note that you and your counterpart will be timed out at the exact same time.

Script [2] for treatments NC_PD and C_PD:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a randomly determined point in time. Any second within the maximum allotted time of 300 seconds (5 minutes) is possible. Note that you will be timed out independently of your counterpart.

Script [3] for treatment CONTROL:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a randomly determined point in time. Any second within the maximum allotted time of 300 seconds (5 minutes) is possible.

Chapter 2:

Third-Party Intervention and Perception Manipulation[†]

[†]I would like to thank Robert Sugden, Anders Poulsen, and Odile Poulsen for financial support and helpful guidance. I would also like to thank Amrish Patel for serving as a discussant at the design stage of the experiment and Silvia Sondereggar for useful comments. Finally I would like to thank the audience of the 2019 CCC (CBESS-CEDEX-CREED) meeting for their feedback.

2.1 Introduction

In *Leviathan* (1651), one of the most influential expositions of *Social Contract Theory*, Thomas Hobbes sees the delegation of human rights to an authority as a requirement to tame opportunistic behaviour and to keep social interactions in order. Without such an authority and its ability to enforce normative behaviour through legal sanctions, humans would live in a state of anarchy characterised by a “war of all against all”.

In modern societies, authorities play an important role in deterring misconduct and criminal behaviours. On the streets, policemen intervene and defuse emerging conflicts amongst citizens. In the courtroom, judges and juries evaluate disputes and impose sanctions in an attempt to restore justice and to deter future misconduct. Deterrence, however, is imperfect because low detection probabilities and mild sanctions may allow law violations to pay in expected terms. In situations outside the reach of the legal system, society relies on the willingness of its citizens to uphold justice and enforce norm compliance on their behalf.

Evidence stemming from the field and the laboratory suggests that people (such as bystanders or acquaintances) are indeed willing to compensate victims and to punish wrongdoers even when they themselves are not affected by the norm violation and despite intervention being personally costly (Fehr and Gächter, 2002; Fehr and Fischbacher, 2004). Because such interventions cannot be explained by models of standard self-interest, they are often claimed to originate from an altruistic concern for the well-being of others and from a preference for norm compliance.

It is now well established in the experimental literature, however, that seemingly altruistic behaviour in a variety of games is often motivated by extrinsic factors such as social pressure or concerns over how actions reflect on one’s social- or self-image (Cain, Dana and Newman, 2014). Generosity in dictator games, for instance, is significantly reduced under conditions which credibly secure anonymity (Hoffman et al., 1994; Hoffman, McCabe and Smith, 1996). It has also been documented that people are biased in self-serving directions. They, for instance, interpret ambiguity about the consequences of their actions, other people’s intentions, or the resolution of uncertainty in ways which best align with their own self-interest (Dana, Weber and Kuang, 2007; Di Tella et al., 2015; Exley, 2015; Haisley and Weber, 2010). We suspect similar forces to be at work in the domain of third-party intervention.

Previous research has mainly relied on the use of two experimental paradigms to study people’s willingness to intervene altruistically: the *Third-Party Punishment Game* and the *Third-Party Compensation Game* (TPPG and TPCG, respectively). In these games, a “third-party” observes the outcome of a dictator game played between two other players and subsequently decides whether to incur a cost to punish the dictator (in the TPPG) or compensate the recipient (in the TPCG). A common

feature of these games is that the dictator’s actions transparently map into outcomes which means that a third-party can easily assess the degree to which the dictator behaves selfishly and violates the focal norm of distributional equality. This is not necessarily the case in the real world where uncertainty may arise over a perpetrators underlying intentions or the severity of harm inflicted on a victim. For instance, a bystander might witness a person verbally insulting or physically approaching another on the street, not knowing the cause of this action or how much of an impact the insult or the approach has on the victim’s mental or physical well-being. Under uncertainty, the bystander might withhold intervention by telling himself that the offence “probably wasn’t that bad”.¹⁵ Uncertainty is also present in situations where bystanders have to step in to *prevent* a foreseeable norm transgression. Imagine you observe a verbal dispute between two people. You anticipate that one of two things will happen: (i) the parties will resolve the conflict peacefully, or, (ii) the conflict escalates and immediate harm is the consequence. As a bystander, you consider whether or not to intervene. If you prefer to avoid getting involved, you may convince yourself that scenario (i) is most likely to happen.

In this paper, we seek to investigate two questions. Firstly, will third-parties intervene at a cost to *prevent* potential harm? To answer this question, we designed a third-party intervention game which features uncertainty about the existence and severity of norm violation. Secondly, will subjects exploit the existing uncertainty in a self-serving way so as to avoid costly interventions? To the best of our knowledge, our study is the first to investigate the role of self-serving belief formation in altruistic third-party intervention.

The remainder of this paper is structured as follows. Section 2.2 reviews the related literature in more detail. Section 2.3 elaborates on the experimental design, hypotheses, procedures and results of our main experiment. In Section 2.4 we present the design and results of a follow-up experiment. Section 2.5 contains a general discussion. Section 2.6 concludes the analysis.

2.2 Related Literature

In this section, we consecutively review the literatures on *third-party intervention* (Section 2.2.1) and *belief distortion* (Section 2.2.2).

2.2.1 Third-Party Intervention

An intervention more generally refers to an intentional action of becoming involved in a situation with the aim of improving it or preventing it from getting worse.

¹⁵A similar story is told by Bicchieri (2006, p. 182) about bystanders to emergencies being afraid of embarrassing themselves by overreacting. Trying to figure out if there is a cause for concern, they interpret the inaction of others as a sign that “probably there is nothing to worry about”.

The experimental literature has mainly focused on the former by studying people's willingness to *restore* justice following a norm transgression. The most prominent experimental approaches to the study of third-party intervention allow subjects in the lab to either punish a norm transgressor or to compensate the victim of a norm transgression. Because these interventions are privately costly and are carried out by subjects who are not directly affected by the norm transgression, they are often referred to as "altruistic" interventions.

In the well-known Third-Party Punishment Game (TPPG; Fehr and Fischbacher, 2004), for example, three players (A, B, and C) are endowed with an equally sized amount of money. Player A (the dictator) is given the opportunity to transfer an amount of money from B's (the victim's) account to his own. Player C (the third-party) observes A's decision and can decide to punish A by deducting his earnings at a personal cost. In the seminal paper on the TPPG, almost two-thirds of third-parties in the experiment punished violations of the distributional fairness norm and punishment increased the more the norm was violated.

Many studies have enriched the basic game by taking into account features which are frequently present in real-world settings. Nikiforakis and Mitchell (2014) enriched a third-party's choice set by allowing for both punishment and reward.¹⁶ They find that the demand for costly punishment is reduced in presence of reward opportunities. The authors propose an additional rationale for punishment which cannot be explained by preferences over material payoff distributions, namely, the signalling of disapproval. In the presence of reward opportunities, many individuals signal their disapproval by withholding reward.

A relative reluctance to punish has been identified in environments which allowed third-parties the additional opportunity of compensating the victim of a norm transgression. In Chavez and Bicchieri (2013) and Lotz et al. (2011), for instance, subjects restored justice more often through compensation. This is consistent with the "do-no-harm" principle according to which people are reluctant to inflict harm on others (Molenmaker, de Kwaadsteniet and van Dijk, 2016). In a study where punishment decisions were reached either individually or in a group, Molenmaker, de Kwaadsteniet and van Dijk (2016) find that groups are more willing to impose punishment than individuals. The authors argue that the diffusion of responsibility in groups alleviates the restraint that individuals experience when they are solely responsible for the implementation of punishments.

Several studies have questioned the robustness of altruistic punishment and compensation from a methodological viewpoint by claiming that elements of the standard design allow for alternative explanations (see e.g. Pedersen, Kurzban and McCullough, 2013). Jordan, McAuliffe and Rand (2016) address two such concerns,

¹⁶The relevance of choice sets has previously been reported in the domain of dictator game generosity by Bardsley (2008) and List (2007).

namely, that punishment could be motivated by envy (as selfish dictators earn higher payoffs), or could be influenced by the use of the strategy method (as subjects might infer and comply with the experimenter's research hypothesis). None of the manipulation, however, significantly affected punishment. Another concern addresses audience effects. Varying conditions of anonymity, Kurzban, DeScioli and O'Brien (2007) indeed find an effect of norm violation on third-party punishment in Trust and Prisoner's Dilemma games. In the former, e.g., 67% of third-parties punished when other participants could observe the decision, while only 42% did so when anonymity was credibly secured.

Although extrinsic factors are shown to matter, overall the literature supports the idea that third-parties also care intrinsically about the compliance with norms and the well-being of others. One potentially relevant aspect however has yet received little attention in the aforementioned analyses: the role of self-image concerns. Self-image concerns arise from a person's desire to arrive at a positive *self*-assessment when reflecting on one's own actions. An internal tension or discomfort is experienced where these actions mismatch a person's moral ideals. Insufficiently accounted for in previous research is the idea that the cause or severity of an interpersonal conflict may be uncertain and that bystanders who should feel obliged to intervene can exploit the inherent uncertainty in self-serving ways. By holding favourable or motivated beliefs e.g. about the underlying level of aggression of a perpetrator or the harm imposed on a victim, a third-party may reduce his or her efforts in resolving the conflict without suffering any moral discomfort.

2.2.2 Perception Manipulation

Cognitive dissonance theory (Festinger, 1962) proposes that people try to achieve internal consistency between their opinions or beliefs and their actions. A tension or dissonance is experienced where these aspects conflict as is the case for example when someone desires to be selfish but dislikes being perceived in a negative light either by others or by himself. In such situations, people can reduce the experienced tension e.g. by reducing self-interested behaviour or by engaging in self-deception.

A vast body of research has documented that other-regarding behaviour in the lab is reduced when subjects can obfuscate responsibility for outcomes or when being enabled to process information in self-serving ways. To provide a few examples, Konow (2000) and Rodriguez-Lara and Moreno-Garrido (2012) show that subjects who face allocation decisions selectively employ justice principles which best align with their own financial self-interest. Dana, Weber and Kuang (2007) demonstrate that subjects in binary dictator games are more selfish when they can remain wilfully ignorant about the consequences of their actions for their counterpart. Exley (2015) shows that subjects use risk as an excuse not to give to charity. Hamman,

Loewenstein and Weber (2010) argue that individuals use delegation to avoid the discomfort of having to implement selfish decisions themselves. Haisley and Weber (2010) show that subjects use ambiguity in an experimental labor market as an excuse for letting-off workers.

In a study closely related to ours, Di Tella et al. (2015) document that subjects distort their beliefs about others' altruism to justify their own selfish behaviour. In their experiment, subjects are matched in groups of two and are assigned either the "allocator" or the "seller" role. Each player is endowed with 10 tokens. Actions are chosen *simultaneously*. The allocator chooses how many tokens to transfer from the seller to himself. The seller determines the price at which both players can cash their tokens in at the end of the experiment. The seller chooses between (i) £2 per token, or (ii) £1 per token. Moreover, the seller receives an additional side payment of £10 if he chooses the low token value.¹⁷ After decisions are made, allocators are asked to guess the percentage of subjects in the seller role choosing the small token value; this percentage is their main outcome variable. Correct guesses are rewarded with a substantial bonus. There are two conditions. In *able=2*, allocators are constrained to transfer a maximum of 2 tokens to their account. In *able=8*, allocators can transfer up to 8 tokens. Importantly, allocators are treated silently, meaning that the seller is uninformed which constraint applies to a matched allocator. Consequently, the allocator's belief regarding the likelihood that the seller chose the low token value should not differ across conditions. In accordance with their research hypothesis, however, allocators in the *able=8* condition believe that a higher percentage of sellers chose the low token value compared to allocators in the *able=2* condition (69% compared to 49%). Since in *able=8*, greater selfishness is possible, allocators have a higher incentive to manipulate their beliefs about the seller's type to justify taking more tokens; in other words, they are *conveniently upset*.

Although we borrow design features from Di Tella et al. and likewise look at motivated beliefs, the angle from which we approach this topic is different from theirs. Di Tella et al. look at motivated beliefs of *second*-parties who are directly affected (in monetary terms) by the misconduct of their counterpart. On the contrary, we investigate the formation of motivated beliefs by *third*-parties who are financially unaffected. In our setup, the focus is on concerns over perceived morality and we aim to investigate whether third-parties generate convenient beliefs which allow them to withhold costly interventions. Another difference concerns the predicted direction of distorted beliefs: whereas subjects in Di Tella et al. are predicted to generate negative beliefs about a counterpart to justify selfish actions, subjects in our experiment have to generate positive beliefs about the likely behaviour of a potential norm transgressor to justify a lack of intervention.

¹⁷The authors also refer to their game as a "corruption" game, namely a dictator game where the recipient can reduce the size of the pie in exchange for a side payment.

2.3 Experiment 1

We conducted two experiments to explore the role of belief distortions for third-party behaviour. Experiment 1 was designed to provide an environment which offers room for belief distortions to evolve. This experiment however does not allow to provide a causal interpretation of the effect of observed beliefs (and possible distortions thereof) on behaviour for reasons which will be outlined below. Experiment 2 is a complementary study which induces exogenous variation of beliefs, thereby allowing us to obtain a better understanding of the relationship between beliefs and behaviour in our experiment.

2.3.1 Design

We introduce a modification of the conventional third-party intervention paradigms (TPPG and TPCG; see introduction) which features uncertainty about the existence and severity of norm violation.

As in the conventional games, there are three roles in our game: Player A (the dictator), Player B (the victim), and Player C (the third-party). Players A and B start with an equal endowment of 10 tokens in their accounts. These tokens are worth £0.80 per unit. Player A has a chance of transferring tokens from passive Player B’s account to his own by *claiming* tokens from Player B’s account. We denote Player A’s claim by $c \in \{0, 1, 2, \dots, 10\}$. The actual transfer however can differ from Player A’s claim depending on the behaviour of Player C.

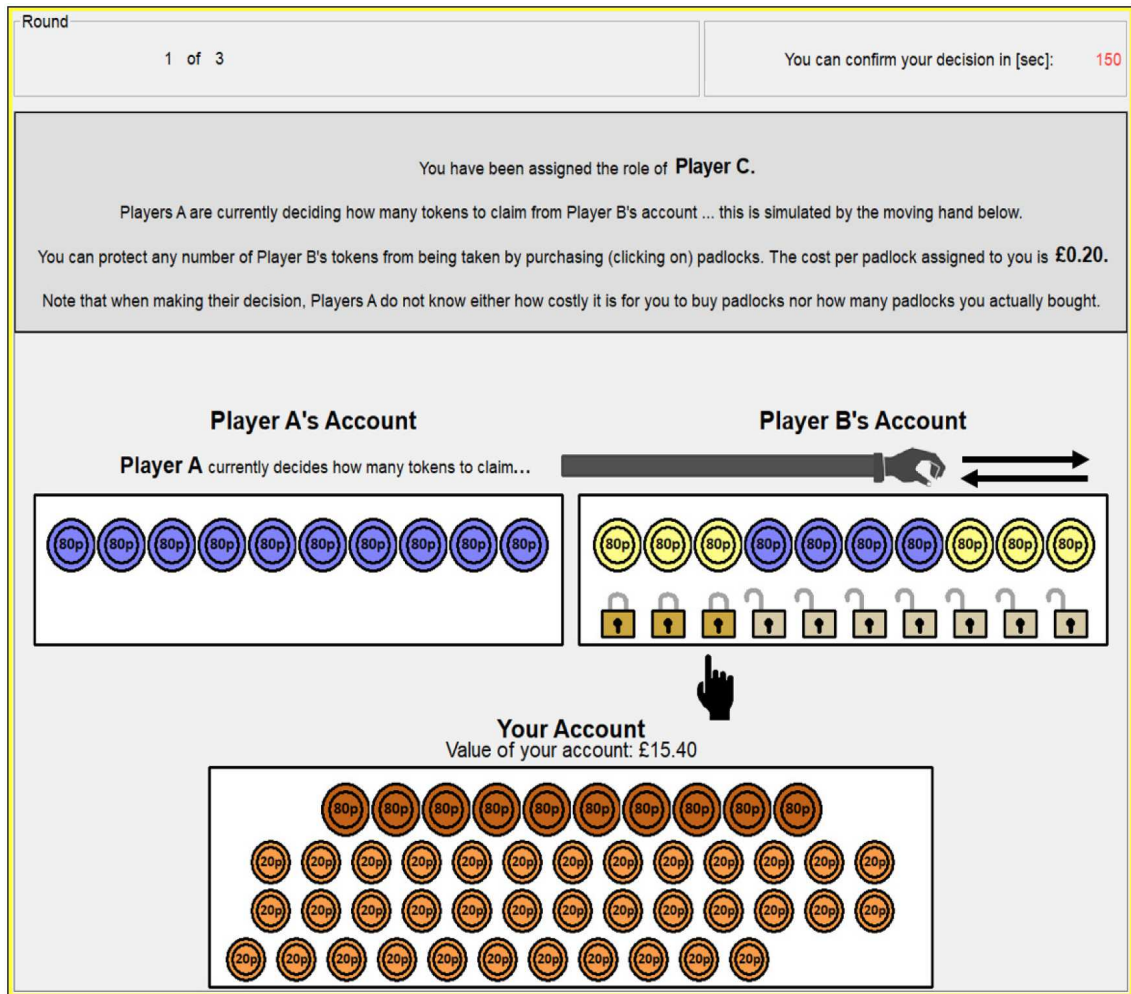
Player C (whose behaviour we are ultimately interested in) is endowed with tokens worth £16. By sacrificing some of her own tokens at a cost, Player C can protect any number of Player B’s tokens from being transferred. We denote Player C’s decision of how many of Player B’s tokens to protect by $p \in \{0, 1, 2, \dots, 10\}$. The transfer, denoted by t , is determined by the function $t = \max\{0, c - p\}$ which means that tokens which are claimed but not protected are transferred from Player B’s to Player A’s account.

Uncertainty is introduced by letting Players A and C decide simultaneously.¹⁸ This means that Player C decides how many tokens to protect not knowing how many tokens Player A will claim. Similarly, Player A decides how many tokens to claim not knowing how many tokens Player C will protect.

Since a prerequisite of testing the strength of moral norms is to create an environment where these norms are salient, we opted for a rather loaded frame that (i) used terms such as “claim”, “take”, and “protect” in the instructions and (ii) also visualised these concepts graphically. Figure 2.1 illustrates how we implemented the game in the laboratory by providing a screenshot of Player C’s decision interface.

¹⁸Appendix 2.A compares our game with a sequential move game with incomplete information.

Figure 2.1: Player C's Decision Screen



Note: Arrows were not part of the original screen and were added to illustrate motion.

Player A was in control of the hand above Player B's account which allowed him to indicate his token claim. By selecting one of the tokens of Player B's account, that token and all tokens to its left would be claimed. Claimed tokens were flagged in a player-specific blue colour. Recall that due to the simultaneous move structure of our game, it was impossible for Player C to know what the actual token claim was until the end of the experiment. In order to make the simultaneous move structure and the uncertainty surrounding players' decisions more comprehensible to subjects, we simulated the choice process of a counterpart on players' screens. In the case of Player C, this meant that the hand indicating the claim would move up and down along Player B's account whilst Player C was deciding how many tokens to protect.¹⁹

Player C could indicate his protection decision by activating padlocks of Player B's account. Selecting a padlock meant that the respective padlock together with all padlocks to its left would activate and click into place. Activated padlocks

¹⁹Similarly so, Player A's decision screen simulated Player C's decision process and Player B's waiting screen simulated both Player A's and Player B's decision processes. Screenshots of Player A's and B's decision interfaces can be found in Appendix 2.B.

prevented tokens above them from being claimed. Whilst activating padlocks, Player C saw some of his own tokens disappear as a consequence of protection being costly. Choices however were non-binding until a minimum of 150 seconds elapsed and a button appeared on screen allowing subjects to confirm their final decision. We thought that a minimum time requirement would increase the chances that subjects would engage with the task and as a consequence form beliefs about the likely behaviour of their counterpart. These beliefs are the focus of our investigation.

The described game was repeated over three rounds. After each round, subjects were re-matched ensuring that no subject would ever interact with another one for more than one round. At the same time, we rotated their roles such that, by the end of the experiment, every subject had played once as Player A, once as Player B, and once as Player C. The roles were rotated in a clockwise order such that a third of our subjects played the order $A \rightarrow B \rightarrow C$, another third played $B \rightarrow C \rightarrow A$, and the remaining third played $C \rightarrow A \rightarrow B$. Subjects were informed that only one of the three rounds would randomly be selected at the end of the experiment to be payoff relevant. This feature of our design allowed us to increase the number of data points collected for C-Players whilst preserving the one-shot nature of interactions.

The treatment assigned was a variation of Player C's protection cost. In a low cost condition (L-Cost), the cost of protecting Player B's tokens was merely £0.20 per unit whereas in a high cost condition (H-Cost) the cost was quadrupled to £0.80 per unit. It was common knowledge to all subjects that for any given round, the computer would independently assign one of the two cost conditions with equal probability to the C-Player of the respective round. Importantly, only Player C herself learned which cost condition applied to her in a given round. The privacy of this information was moreover maintained beyond the concluding outcome stage of the experiment where subjects were merely informed about the claim and protection rates that applied in the payoff relevant round, not however about the assigned cost condition or the final payoffs of the selected C-Player. We further elaborate on the significance of this "silent" treatment design in the hypotheses section of our paper.

For every round of the experiment, we elicited Player C's beliefs about the claim following his protection decision. Figure 2.2 illustrates what C-Players saw on their screen. We asked them to consider all A-Players of a given round (excluding the one of their own group) and to guess in which bracket (out of 10) the average claim of these A-Players would fall.²⁰ We incentivised the accuracy of beliefs by rewarding correct guesses with a non-negligible bonus of £5 which was added to subjects' final earnings at the end of the experiment.

²⁰We asked not to consider the A-Player of one's own group to align the current design with that of Experiment 2 where we elicited protection choices using a strategy method approach. This exclusion was necessary to preserve the uncertainty about a counterpart's actual behaviour in that experiment. More details will follow in Section 2.4.1.

Figure 2.2: Belief Elicitation Screen

Round
1 of 3

Bonus Question: for an additional £5.

In the current round, 6 participants in the lab played in the role of Player A. Excluding the Player A of your own group and only considering the other 5 participants who played in the role of Player A in groups other than yours:

How many (out of 10) tokens do you think have these 5 participants claimed on average from Player B's account?

Please indicate your guess by ticking one (and only one) of the categories below.

If you correctly guess the average claim, you will receive an additional £5 as a price at the end of the experiment.

- The 5 Players A claimed on average less than 1 (out of 10) tokens
- The 5 Players A claimed on average at least 1 but less than 2 (out of 10) tokens
- The 5 Players A claimed on average at least 2 but less than 3 (out of 10) tokens
- The 5 Players A claimed on average at least 3 but less than 4 (out of 10) tokens
- The 5 Players A claimed on average at least 4 but less than 5 (out of 10) tokens
- The 5 Players A claimed on average at least 5 but less than 6 (out of 10) tokens
- The 5 Players A claimed on average at least 6 but less than 7 (out of 10) tokens
- The 5 Players A claimed on average at least 7 but less than 8 (out of 10) tokens
- The 5 Players A claimed on average at least 8 but less than 9 (out of 10) tokens
- The 5 Players A claimed on average at least 9 or more (out of 10) tokens

You will be informed whether or not you guessed correctly at the end of the experiment!

2.3.2 Hypotheses

The core hypothesis our experiment was designed to test is that C-Players form motivated beliefs about the type of a matched A-Player to reduce the level of costly protection they feel obliged to provide. Recall that C-Players were treated silently, i.e. their assigned cost condition was private information. Since A-Players never learn which cost condition applied to a matched C-Player in a given round, A-Player behaviour cannot depend on the assigned cost condition. Under the assumption of rationality in C-Players' beliefs about the behaviour of A-Players, C-Players' beliefs should not depend on the assigned cost condition either. Our null hypothesis therefore states:

Hyp. 0: *C-Players' beliefs about the claim are independent of whether the treatment is L-Cost or H-Cost.*

If however, C-Players looked for ways to avoid costly protection, without having to suffer any moral discomfort, they could convince themselves that less protection is needed. The alternative hypothesis is about non-rationality in C-Players' beliefs:

Hyp. 1: *Beliefs about the claim are lower under H-Cost than L-Cost.*

Our high cost condition was calibrated with the aim of providing sufficient incentives for belief distortions to evolve. Notice that we contrast this condition with an alternative where protection is extremely cheap; full protection can already be implemented at a small cost of £2 out of £16. The idea is that subjects in the low cost condition face less of an incentive to distort their beliefs since the monetary benefit of doing so (evading protection costs) is rather small. Consequently, we expect beliefs under this condition to be closer to subjects' true (or, undistorted) beliefs.

We assume that subjects distort their beliefs to evade a perceived obligation to protect.²¹ Under this interpretation, we would expect fewer tokens to be protected under H-Cost than L-Cost.

Hyp. 2: *C-Players protect fewer tokens under H-Cost than L-Cost.*

Our discussed belief channel is not the only possible explanation for differences in observed protection. First of all, the cost variation itself is expected to affect protection. Second of all, the cost variation changes the relative payoff distributions that can be obtained between treatments as a result of the protection decision. In Section 2.4 we present a follow-up experiment which induces exogenous variation of beliefs and which allows us to control for these alternative explanations.

2.3.3 Procedures

The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted in the *Laboratory for Economic and Decision Research* (LEDR) at the University of East Anglia. A total of 144 participants recruited from the local student population took part in the study. We conducted 8 sessions in the Autumn of 2018, each of which lasting around 50 minutes. 18 participants took part in each session. Average earnings were £13.70, with a minimum of £3 and a maximum of £24 (including a £3 participation fee).

Upon arrival, participants were randomly assigned to computer terminals by drawing their desk number. Each computer was located in a separate cubicle which inhibited visual interaction or communication. Anonymity amongst participants was secured because at no point during or after the experiment did any participant receive identifying information about his or her peers. We also took great care in the instructions emphasising that the experimenter would not be able to link the generated data to any participant as a person. Participants received a hard copy of the instructions and were asked to follow along as the experimenter read

²¹Belief distortions can also occur for non-instrumental reasons. We address this possibility in Section 2.5.

the instructions out aloud. The instructions included hands-on exercises meant to familiarise subjects with the stages, screens, and mechanisms of the following experiment. Clarifications were provided on an individual basis. Participants were asked to answer a set of five control questions following the instructions. The experiment concluded with a brief questionnaire asking for demographic information and an assessment of the difficulty of the experimental tasks. Privacy was guaranteed during the payment phase by asking participants to individually collect their final earnings from an experimental assistant at the end of the experiment.

2.3.4 Results

All data referred to in this section is also subsumed in Table 2.1 which provides summary statistics on various outcomes of our experiment, all broken down by treatment and role order.

2.3.4.1 Protection Behaviour

Figure 2.3 depicts a breakdown of average protection behaviour in our experiment. We find that irrespective of the order of roles by which subjects encountered the

Figure 2.3: Average Protection Behaviour by Treatment and Role Order

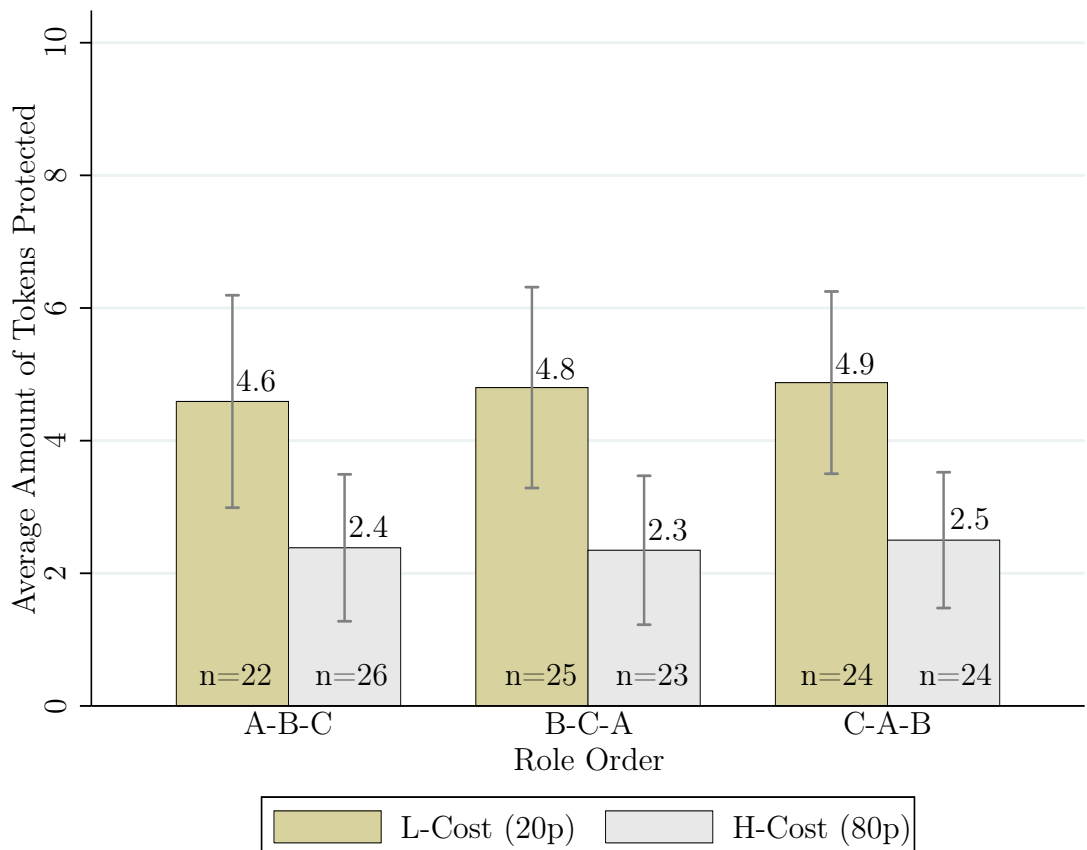


Table 2.1: Summary Statistics for Experiment 1

	n	Average Protection	Average Claim	Average Belief	Spearman Prot.\Belief
Experiment 1	144	3.6	8.3	7.3	0.28***
H-Cost (£0.80)	73	2.4	8.4	7.1	0.24**
A-B-C	26	2.4	8.8	7.3	0.17
B-C-A	23	2.3	7.9	6.8	0.25
C-A-B	24	2.5	8.3	7.1	0.34
L-Cost (£0.20)	71	4.8	8.3	7.5	0.29**
A-B-C	22	4.6	8.8	7.5	0.27
B-C-A	25	4.8	8.0	7.3	0.34*
C-A-B	24	4.9	8.0	7.6	0.23

Note: Rows 4-6 and 8-10 break the data down by the order of roles that subjects encountered in the experiment. Columns 3 and 4 report the average number of tokens protected or claimed, respectively. Column 5 relates to the belief band that C-Players expected the average out-group claim would fall in. Column 6 reports the spearman correlation coefficient between the number of tokens protected and C-Players' beliefs.

three rounds of the experiment, subjects protected fewer tokens when protection was relatively more expensive. Pooling across role orders, the average amount of tokens protected under L-Cost is twice that of H-Cost (4.8 vs. 2.4; $Z = 3.744$, $p < 0.01$, one-tailed), thereby supporting hypothesis 2.

Result 1. *C-Players protect fewer tokens under H-Cost than L-Cost.*

It is reasonable to expect C-Players to protect less as a consequence of protection being more costly. For subjects with moral concerns however, protecting less could still be quite costly as such subjects may experience a disutility from acting against what they think is the morally correct action. One way of avoiding both types of costs (i.e., pecuniary and moral) is to convince oneself that A-Players will claim fewer tokens.

2.3.4.2 Beliefs about the Claim

Before testing our main hypothesis of distorted beliefs, it is worthwhile to look at the relationship between reported beliefs and protection behaviour more generally.

Figure 2.4: Beliefs and Protection Behaviour by Treatment

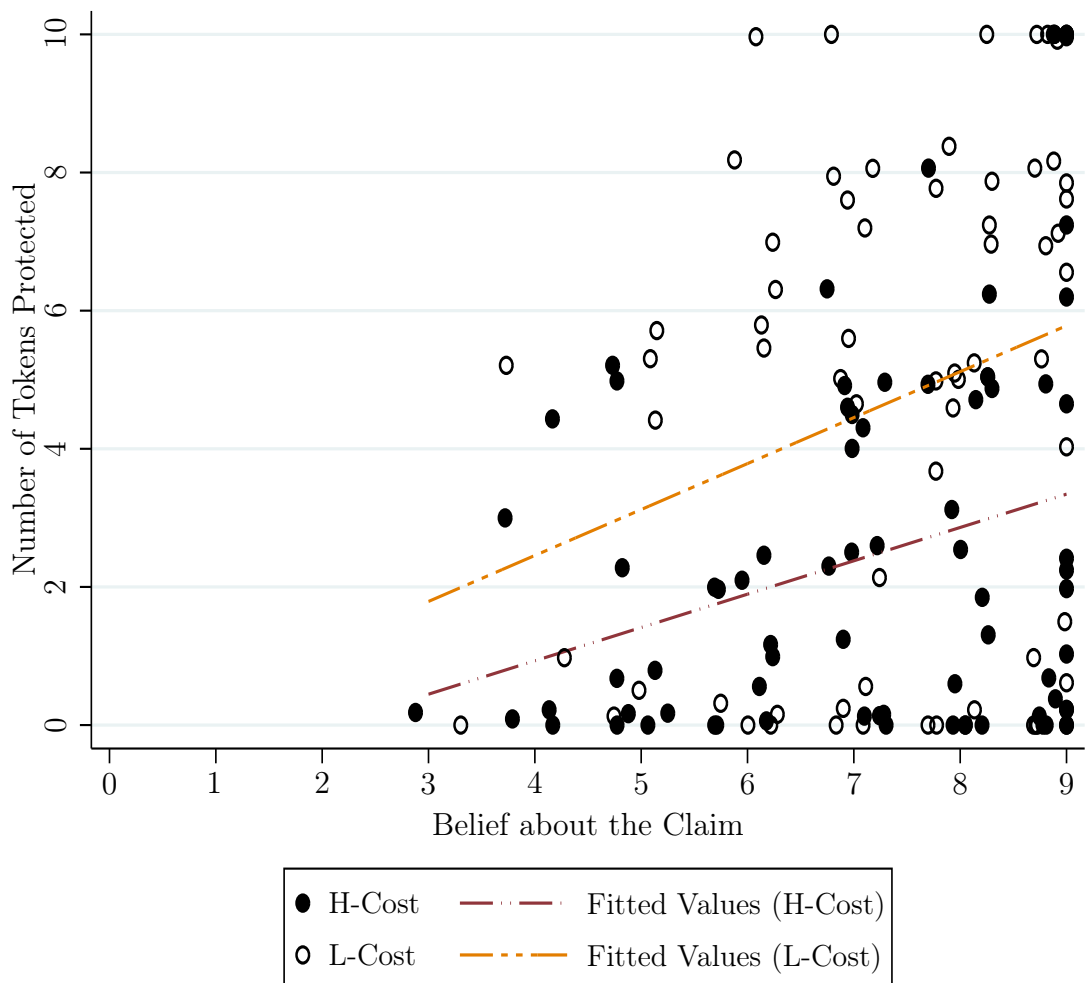
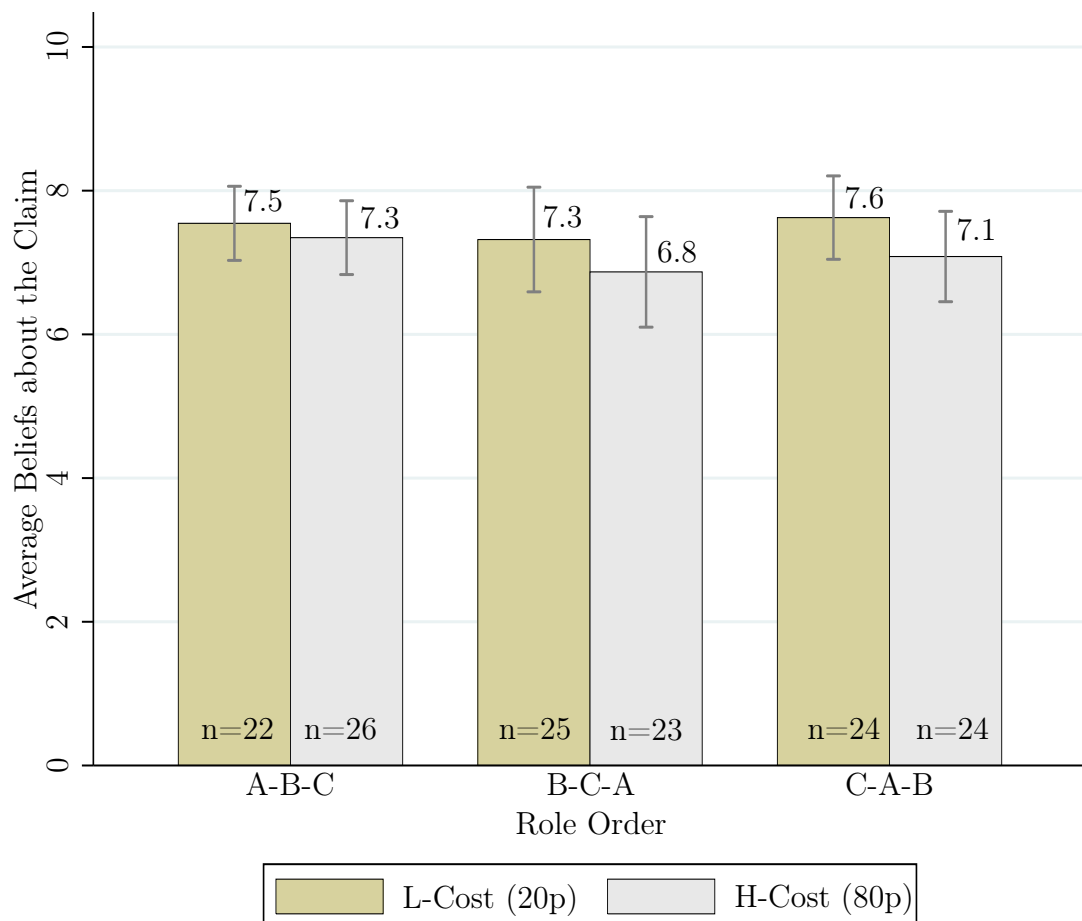


Figure 2.4 depicts the associated scatter plots for each cost condition. The x-axis refers to the belief category (out of 10) chosen by a subject, whereby “0” indicates the lowest category (average claim less than 1) and “9” indicates the highest category (average claim of 9 or more). As is evident from the graph, reported beliefs about the claim correlate positively with the number of tokens protected by a subject under both cost conditions (Spearman correlation coefficient = 0.28, $p < 0.01$, for the pooled sample). While we acknowledge that such a correlation is insufficient to prove a causal relationship between beliefs and behaviour for reasons such as reverse causality, it is at least compatible with the idea that subjects in our experiment acted on their beliefs. This premise gives rise to the idea that subjects could have distorted their beliefs instrumentally, to reduce the level of protection they felt obliged to provide.

To look for evidence of belief distortions, we first inspected reported belief averages which are broken down by treatment and role order in Figure 2.5. What we find is that for all role orders, average beliefs about the claim are lower in the high cost compared to the low cost condition. Belief differences also appear

Figure 2.5: Average Beliefs by Treatment and Role Order



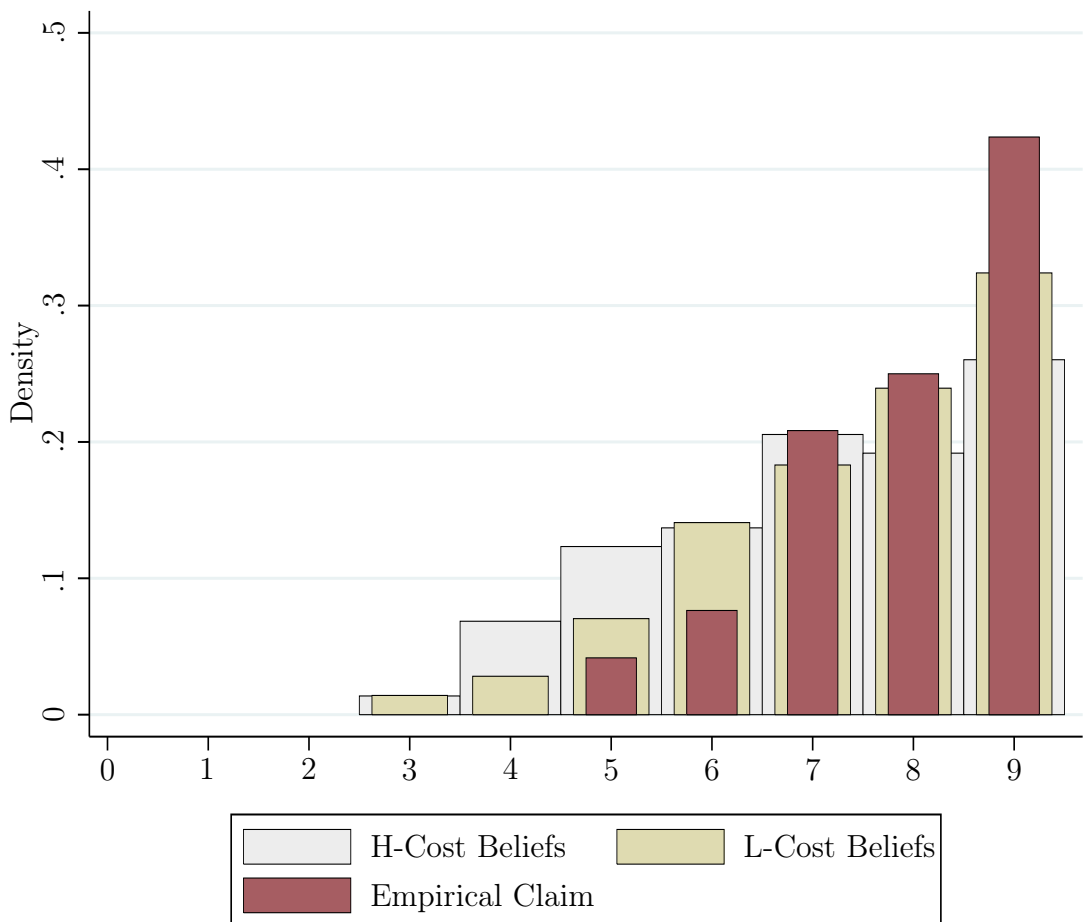
more pronounced for C-Players who had no prior experience as A-Players. However, according to Kruskal-Wallis tests, reported beliefs are statistically indistinguishable across the three role orders (L-Cost: $\chi^2 = 0.167, p = 0.915$; H-Cost: $\chi^2 = 0.532, p = 0.760$). This justifies pooling the data across the role order dimension.

For the pooled data, Figure 2.6 depicts the distributions of beliefs (about the average out-group claim) for each cost condition separately and also contrasts them with the distribution of empirical out-group claims. What we find is a mildly significant difference in beliefs about the claim between our cost conditions in the direction predicted by hypothesis 1 (mean: 7.1 vs. 7.5; ranksum test, $Z = 1.434, p = 0.076$, one-tailed).

Result 2. *There is mild evidence of lower beliefs about the claim under H-Cost compared to L-Cost. This finding is consistent with the idea that subjects entertain motivated beliefs to evade the costs of protection.*

It is also interesting to see that subjects – more generally – appear to hold too optimistic beliefs about the generosity of A-Players. Even when taking condition

Figure 2.6: Distributions of Beliefs and Empirical Claims



L-Cost as the baseline (where we assumed incentives for belief distortions to be small) we find that beliefs about the average claim are significantly smaller as compared to the empirical claims (mean: 7.5 vs. 8.3; ranksum test, $Z = -2.071, p = 0.038$).

Result 3. *We find that irrespective of incentives, C-Players underestimate A-Player opportunism.*

One question which the current experiment is unable to answer is how much of an impact on behaviour the identified belief differences had. Such an identification is confounded under the current design by the simultaneous variation of protection costs and beliefs as well as the endogenous nature of reported beliefs. In the following experiment, we induce exogenous variation of beliefs in an otherwise similar experimental environment to obtain a better understanding of the likely impact that the observed belief differences had on third-party behaviour in our experimental environment.

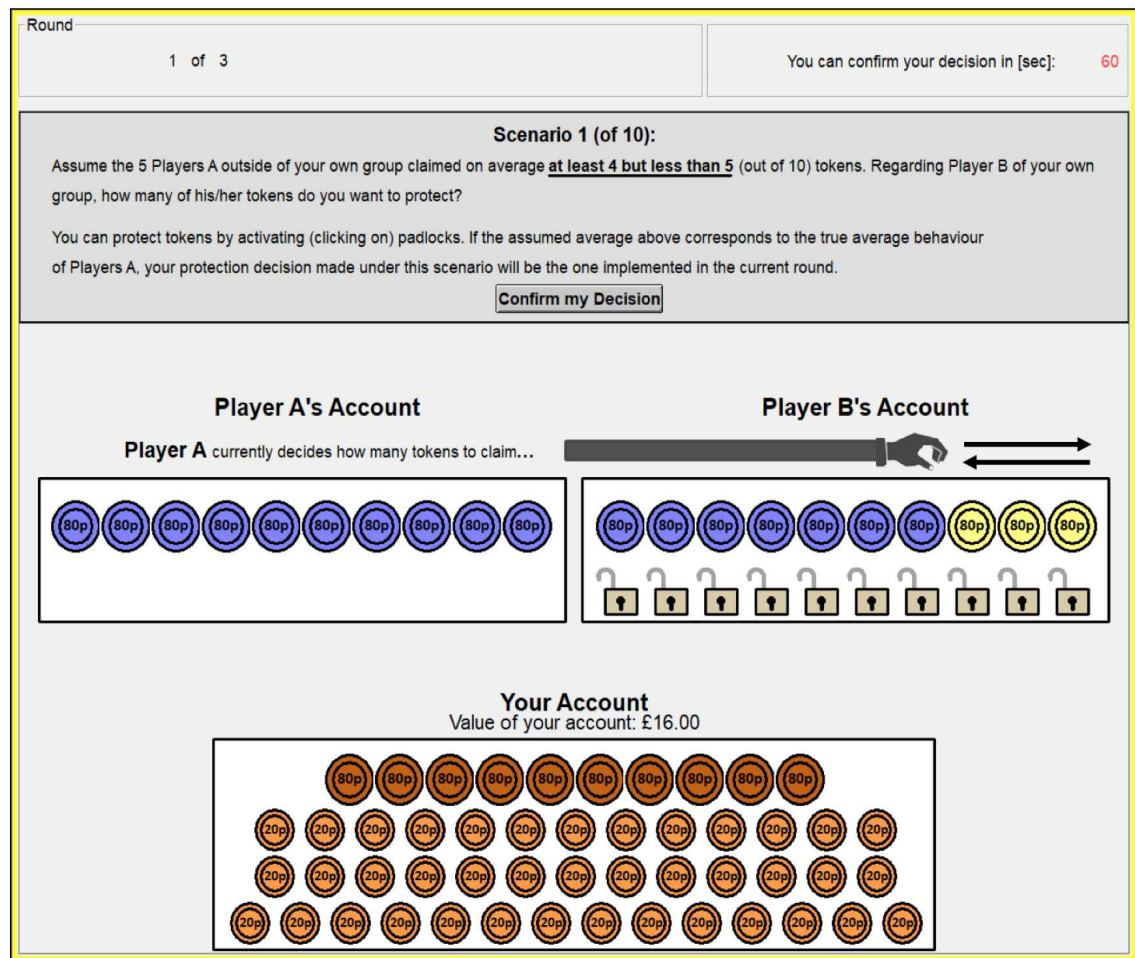
2.4 Experiment 2

2.4.1 Design

The design of experiment 2 closely resembled that of experiment 1, apart from two important changes. Firstly, there was no cost variation, the cost of protection was equivalent to the H-Cost scenario of the previous experiment, i.e. 80p per padlock. Secondly, instead of letting subjects form their beliefs about the claim endogenously as in experiment 1, here we asked C-Players to condition their protection decision on a total of 10 possible scenarios. Each of these scenarios covered one of the 10 belief bands about the average out-group claim used to elicit beliefs in experiment 1. Subjects were asked to submit one protection decision under each scenario and were told that if a respective round was chosen to be payoff relevant, the experimenter would implement the protection decision made under the true scenario, i.e. the scenario that corresponded to the true average out-group claim.

Figure 2.7 provides a screenshot of a subject's decision interface for one of the scenarios. To avoid anchoring effects, we randomised the order by which belief bands mapped onto the sequence of scenarios. As in the previous experiment, there was a minimum time requirement on each screen which however decreased as subjects moved through the scenarios. We also took great care in the instructions and the preceding practice stage in familiarising subjects with the conditional nature of decisions under the current design. A set of control questions was used to check that subjects understood the details of the experiment. The instructions for experiment 2 are provided in Appendix 2.B.4.

Figure 2.7: Player C's Decision Screen for Scenario 1



Note: Arrows were not part of the original screen and were added to illustrate motion.

2.4.2 Procedures

The procedures closely followed those of experiment 1. Since we generate richer C-Player data under the current within-subject design, we recruited a smaller sample of 54 subjects. We conducted 3 sessions with 18 subjects each in the Spring of 2019. Sessions lasted around 50 minutes. Average earnings were £11.5, with a minimum of £3 and a maximum of £19 (including a £3 participation fee).

2.4.3 Results

Table 2.2 provides summary statistics on various outcomes of experiment 2, broken down by the order of roles. The first thing to notice is that the reported summary statistics seem to resemble those obtained under the H-Cost condition of experiment 1 quite closely. Evidence of role order effects is small, with the only apparent difference (again) being that subjects who start off as A-Players in round 1 of the experiment claim relatively more tokens. None of the reported differences in behaviour across the role order dimension however reach statistical significance.

Table 2.2: Summary Statistics for Experiment 2

	n	Average Protection	Average Claim	Spearman Prot.\Belief
Pooled	54	2.6	8.4	0.34***
A-B-C	18	2.8	8.8	0.38***
B-C-A	18	2.5	8.4	0.30***
C-A-B	18	2.6	8.0	0.37***

Note: Rows 3-5 break the data down by the order of roles that subjects encountered in the experiment. Columns 3 and 4 report the average number of tokens protected or claimed, respectively. Column 5 reports the spearman correlation coefficient between the number of tokens protected and the belief bands shown.

Table 2.3: Random Effects Tobit Model Estimations

Model:	RE Tobit	RE Tobit	RE Tobit
Dependent Variable:	Tokens Protected	Tokens Protected	Tokens Protected
Scenario	0.362*** (0.092)	0.441*** (0.050)	0.441*** (0.050)
Role Order			
B-C-A	-0.986 (0.756)	-0.550 (1.017)	
A-B-C	-0.607 (0.719)	0.069 (0.913)	
Scenario \times Role Order			
B-C-A	0.096 (0.129)		
A-B-C	0.147 (0.122)		
Constant	0.479 (0.481)	0.115 (0.547)	-0.040 (0.387)
n	540	540	540

Note: *Scenario* is the belief band for which a certain protection decision was elicited. Standard errors are reported in parentheses. (**, ***): coefficient significantly different from zero at the 10% (5%, 1%) level.

The advantage of experiment 2 is that it induces exogenous variation of beliefs about the claim by asking C-Players to condition their protection decision on a vector of possible claims. To identify the effect of beliefs on protection behaviour we ran random effects tobit model estimations which take into account the panel structure of our data and the censoring of our dependent variable. The results are provided in Table 2.3. As a robustness check, we included two specifications which allow the effect of beliefs on protection behaviour to vary by the order of roles. As already suggested by the summary statistics however, we again find no evidence of role order effects in our experiment. Arriving at our final specification we find that an increase in the belief scenario by 1 belief band, ceteris paribus, is predicted to increase the number of tokens protected by 0.44 tokens.

2.5 General Discussion

The objective of experiment 2 was to isolate the causal effect of belief variations on protection behaviour and – given the similarity of designs – to thereby provide us with an estimate of the likely role that belief variations must have played in experiment 1. Before we proceed with a discussion of our results, we want to briefly address some of the potential weaknesses of our experimental design.

One might criticise our design on the grounds that the strategy method used in experiment 2 makes it more susceptible to experimenter demand effects. This could result in an overstatement of the effects of beliefs on behaviour. While we acknowledge that this is a valid concern, a survey of the literature comparing the strategy method to the direct response method suggests that both typically yield similar results (Brandts and Charness, 2011). Moreover, if the strategy method used in experiment 2 induced subjects to be more responsive to our treatment, we would expect a stronger correlation between beliefs and protection compared to experiment 1. This however was not the case as can be shown by a comparison of spearman rank correlations for the range of beliefs with common empirical support in both studies ($r_s = 0.237$ vs. $r_s = 0.241$)²²; in fact, the correlations are very similar suggesting that our estimates have not been biased by an experimenter demand effect.

Another potential issue concerns the timing of our belief elicitation stage which took place *after* protection decisions had been made. Our decision to let subjects engage with the decision environment first was motivated by the assumption that convenient beliefs would more readily be formed when subjects were given enough time to experience the potential consequences of their actions. Recall that we used a rather loaded frame which also simulated the possible choice outcomes of a counterpart. Moreover, we implemented a minimum time requirement to further increase the chances that subjects would engage with the decision environment and as a consequence form beliefs about the likely behaviour of their counterpart. Although it is true that reported beliefs in our experiment could be the result of subjects adapting convenient beliefs to justify their choices ex-post, we believe such a strategy by which beliefs are distorted for non-instrumental reasons to be very costly in light of the high incentives that we provide for accurate beliefs.

Turning to a discussion of our results, we can connect the findings obtained across the two experiments to obtain an estimate of the likely role that belief differences played in our study. In experiment 1 we found that on average C-Players reduced their protection by $(4.8 - 2.4 =) 2.4$ tokens due to the cost increase. This reduction was accompanied by a mildly significant distortion of beliefs in the magnitude of $(7.5 - 7.1 =) 0.4$ tokens. Under the assumption that subjects distorted their beliefs

²²Whereas experiment 2 by design generates a belief distribution with full support, in experiment 1 no subject reported a belief about the claim of less than 3 tokens.

instrumentally – i.e. to justify protecting less – the results of experiment 2 suggest that the belief distortions identified in experiment 1 are likely to explain as little as $(\frac{0.4 \times 0.44}{2.4} =)$ 7.3% of the decrease in the number of tokens protected. Instead, what appears to have mattered most is the material cost of providing protection.

Although evidence of strategically distorted beliefs is weak in our experiment, an interesting finding is that subjects more generally held too optimistic beliefs about the generosity of A-Players. As a consequence, de-biasing third-parties e.g. through the transmission of empirical information on the severity of offences may be a desirable policy intervention to increase third-party involvement.²³

2.6 Conclusion

In situations outside the reach of the legal system, society often relies on the willingness of its citizens to uphold justice and to enforce norm compliance on their behalf. A large body of literature has documented that third-parties who are not directly affected by a norm violation are nonetheless willing to intervene at a personal cost to secure or restore justice. The aim of our paper was to investigate whether such “altruistic” interventions would also survive in environments which allow to morally excuse non-intervention. More specifically, we introduced uncertainty in a third-party protection game and hypothesised that subjects would – as a consequence of protection being costly – entertain motivated beliefs allowing them to evade a moral obligation to protect.

Not very surprisingly, we observed that subjects protected fewer tokens when the cost of protection increased. We found that this reduction was accompanied by a reduction of subjects’ beliefs about the claim which is consistent with the idea that subjects entertained motivated beliefs to evade a moral obligation to protect. The discovered belief differences however merely reached marginal significance. Moreover, the relevance of belief distortions for behaviour in our experiment appeared to be small; our estimates suggest that merely 7.3% of the variation in protection behaviour could potentially be attributed to the distortion of beliefs. Instead, what appears to have mattered most is the material cost of providing protection.

Despite the weak influence of strategically distorted beliefs in our experiment, we observe support of subjects exhibiting a general bias of perceiving dictators as more generous than they really are. In light of this finding, we think that the societal transmission of information e.g. about the severity of various civil offences could help de-bias the general population and thereby increase their likelihood to intervene as third-parties confronted with immoral behaviour in the field.

²³It is worth noting that our policy implication contrasts with the more familiar nudging argument (e.g. in relation to organ donation, recycling, tax compliance) that people underestimate norm compliance and need to be told how pro-social the average person is (see e.g. Thaler and Sunstein, 2009). The recent norm-nudge literature is discussed by Bicchieri and Dimant (2019).

Appendix for Chapter 2

2.A Sequential Game Comparison

We acknowledge that our game could have also been modelled differently, e.g. as a sequential game with incomplete information regarding the type of Player A. As an example, consider a game where nature privately assigns one of two types to Player A: either (i) Player A is free to claim any number of tokens from Player B (in which case protection might become needed), or (ii) Player A is unable to claim any tokens from Player B (in which case protection becomes redundant). In the sequential version, Player C moves first knowing that two types of A-Players exist, not however the probabilities with which types are assigned. A procedure akin to Weber and Haisley (2010) could be used who let subjects know that the true probability is drawn from a uniformly distributed set of different probabilities. In this game, C-Players would be asked whether they believed to be matched with an “impaired” or “active” A-Player. The higher the cost of protection, the more C-Players may adopt the convenient belief of impairment which would make protection redundant.

Under such a design however, C-Players would form beliefs about a parameter of the experiment, with very little information to base their beliefs on. We prefer our design on the grounds that beliefs are formed about the *behaviour* of A-Players; C-Players know the problem that A-Players face and the subject pool from which they were recruited. C-Players are therefore able to make intelligent guesses about the behaviour of A-Players even without actually knowing the probability.

2.B Instructions and Screens

2.B.1 Experiment 1 Instructions

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to

you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

In this experiment, a task will be performed for three rounds. At the beginning of the first round, you will randomly be assigned one of three possible roles: Player A, Player B, or Player C. You will then be allocated to a group which includes one Player A, one Player B, and one Player C.

At the beginning of each of the following two rounds, your group and your role in the experiment will change. You will be assigned to an entirely new group of participants in each round which means that you will never be matched with any other participant in the room for more than one round. Moreover, you will be assigned a different role in each of the three rounds. At the end of the experiment, every participant will have played once as Player A, once as Player B, and once as Player C.

In every round, you will start with a freshly generated amount of ‘tokens’ in your private account. Depending on your decisions and/or the decisions of your group members, the amount of tokens in your account can change. Every token has a specific monetary value attached to it.

At the end of the final round, the computer will randomly select *one* of the three rounds of the experiment to determine your final earnings. Your earnings will then be equal to the value of all the tokens which you hold in your private account at the end of the selected round.

The Task

In every round, players start with an endowment of tokens. The total value of a player’s initial token endowment is given as follows:

Player A: £8

Player B: £8

Player C: £16

In the task, Player A will have a chance to claim tokens from Player B’s account and Player C will have a chance to protect Player B’s tokens from being claimed. When we say claim, we mean trying to take. Tokens which are claimed but not protected will be transferred to Player A’s account at the end of a given round.

Please have a look on screen where you can see how to claim and protect tokens. You will be told by the experimenter when to return to the paper instructions.

Start of on-screen practice

Practicing how to claim tokens (1 of 2):

On the screen, you can see the tokens that Player A and Player B hold in their accounts.

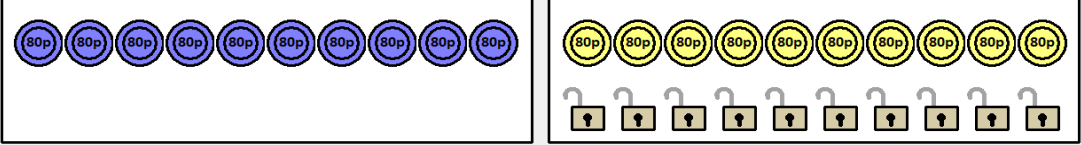
For the moment, please ignore the padlocks that you see on screen, their meaning will be explained soon.

If you are assigned the role of Player A, you will be able to claim tokens from Player B by clicking on them.

Please try to claim 4 tokens by clicking on token number 4 of Player B's account.

Player A's Account

Player B's Account



Practicing how to claim tokens (2 of 2):

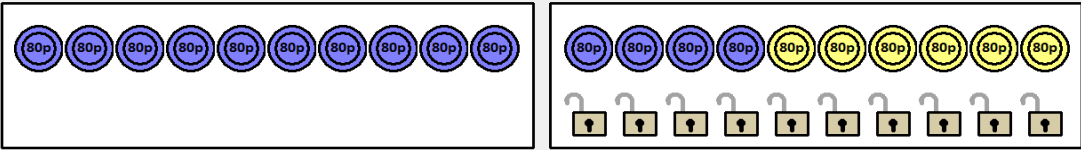
After you made your claim, you can see that a hand appeared to illustrate your claim.

All tokens below and to the left of the hand are currently being claimed by you. Claimed tokens change their colour to blue.

Now, practice a bit by claiming any of the other tokens of Player B. Notice that clicking twice on the same token resets your choice.

Player A's Account

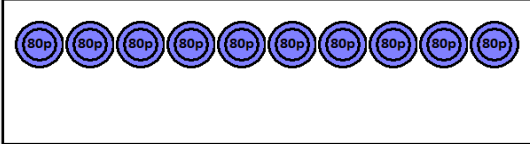
Player B's Account



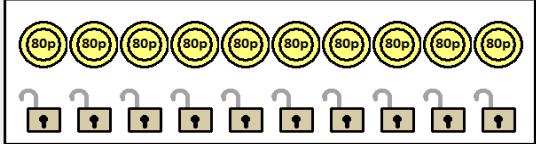
Practicing how to protect tokens (1 of 2):

On the screen, you can see the tokens that Player A, Player B, and Player C hold in their accounts.
 If you are assigned the role of Player C, you will be able to protect Player B's tokens by activating padlocks.
 Activating padlocks is costly for Player C. In this practice stage, the cost is 40p per padlock.
 Please try to activate 6 padlocks by clicking on padlock number 6 of Player B's account.

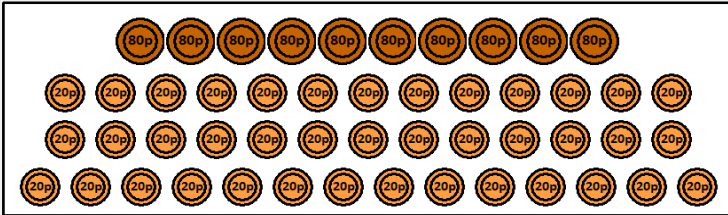
Player A's Tokens



Player B's Tokens



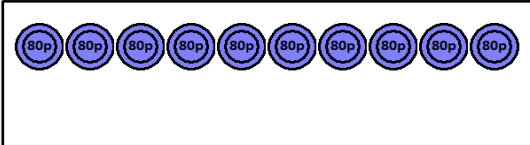
Your Tokens
 Value of your account: £16.00




Practicing how to protect tokens (2 of 2):

After you clicked on a padlock, you can see that a hand appeared to illustrate your decision.
 Activated padlocks click into place and change their colour. All tokens above an activated padlock are currently being protected by you.
 Also notice that the cost of activating padlocks (40p) is deducted from your account in real time.
 Now, practice a bit by activating any of the other padlocks of Player B. Notice that clicking twice on the same padlock resets your choice.

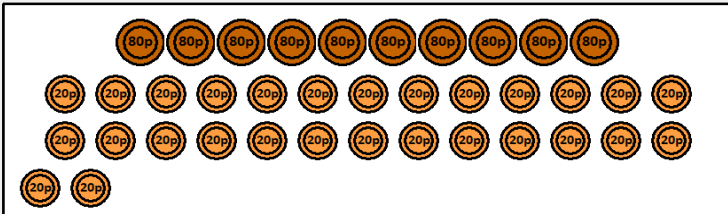
Player A's Tokens



Player B's Tokens



Your Tokens
 Value of your account: £13.60



End of on-screen practice

Different Costs of Protection

The cost of activating a padlock in the main part of the experiment can vary and will either be **20p** or **80p**. In every round, the computer will randomly select one of the two possible costs with equal probability. Only Player C and no other player will ever find out which cost applied in a particular round and for a particular Player C.

Simultaneous Decisions

Player A and Player C will decide *simultaneously*, i.e. at the same time, how many tokens they want to claim or protect. This means that Player A will decide how many of Player B's tokens to claim not knowing how many padlocks Player C will activate. At the same time, Player C will decide how many padlocks to activate not knowing how many tokens Player A will claim.

Determining Payoffs in a Given Round

Tokens which are claimed but not protected are transferred from Player B's to Player A's account. Therefore,

- Player A's payoff in a given round is his initial endowment of £8 plus the value of tokens transferred from Player B's to Player A's account.
- Player B's payoff in a given round is his initial endowment of £8 minus the value of tokens transferred from Player B's to Player A's account.
- Player C's payoff in a given round is his initial endowment of £16 minus the cost of all padlocks he activated.

Notice that you will not receive any feedback on outcomes in any of the three rounds of the experiment until the end of the final round. Remember that one of the three rounds of the experiment will be selected to determine your earnings at the end of the experiment and you will only receive feedback on outcomes and your personal earnings of that specific round.

Bonus: Guessing

In every round of the experiment, *some* of you will have the opportunity to earn additional money by guessing outcomes of the experiment. You will learn more about this during the experiment.

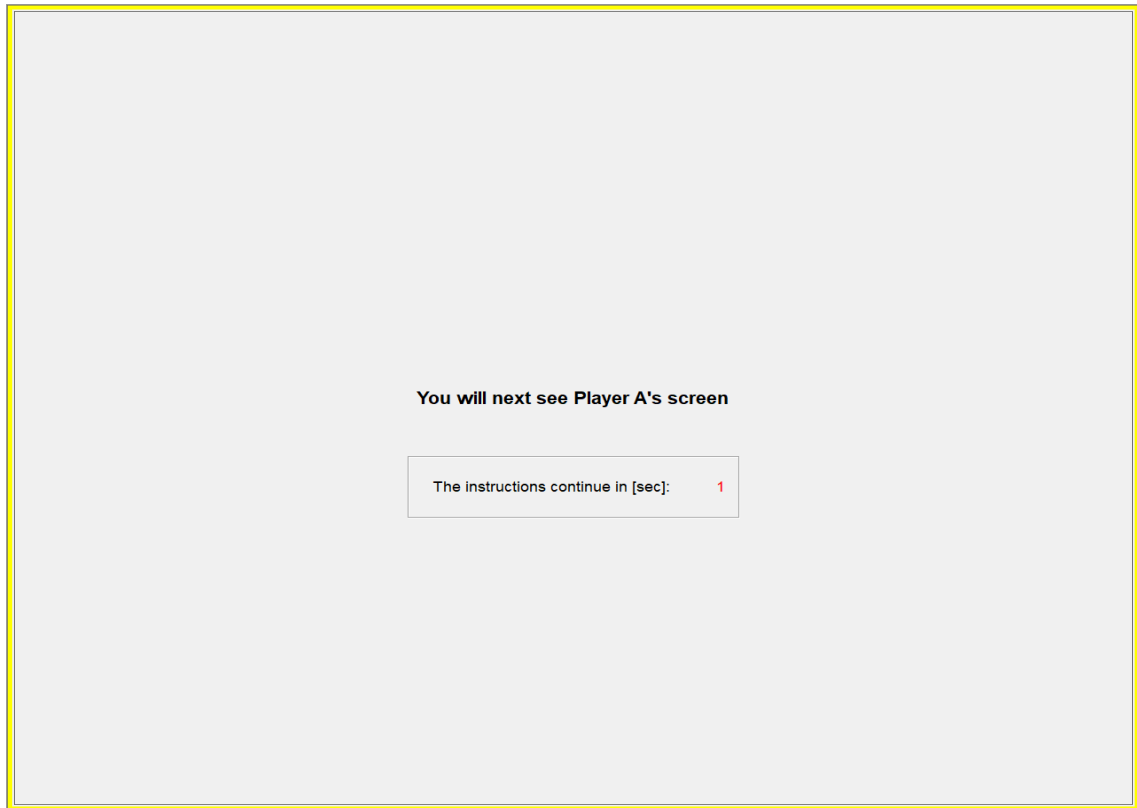
Interface of the Experiment

You already practiced how to claim and protect tokens. We will now take you through the particular screens that you will encounter in the experiment in all three roles to further familiarize you with the interface and processes of the experiment.

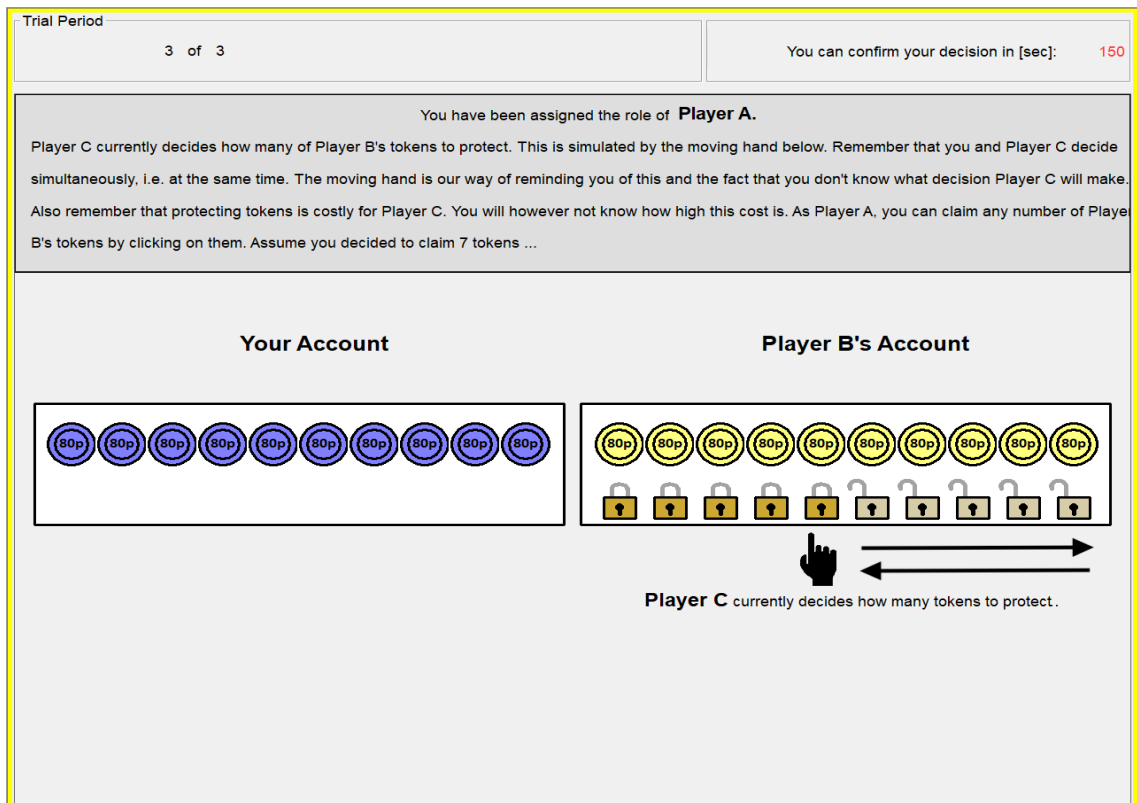
Please follow along on screen.

2.B.2 Example Round Screens

Player A Intro Screen



Player A Screen



Note: Arrows were not part of the original screen and were added to illustrate motion.

Player A Screen (cont.)

Trial Period
3 of 3

You can confirm your decision in [sec]: 150

Your screen will then tell you how many tokens you would obtain for all possible decisions of Player C (which are represented by the moving hand). Remember that tokens which are claimed but not protected will be transferred to your account. Those are the blue tokens that you see on your screen. You will have to spend a minimum time of 150 seconds on this screen before a button appears which will allow you to confirm your final decision. Even after the confirmation button appeared, you can still take as much time as you need and change your choice as often as you like (until you press the button).

Your Account

Player B's Account

The diagram shows two accounts: 'Your Account' and 'Player B's Account'. 'Your Account' contains 10 blue tokens, each labeled '80p'. 'Player B's Account' contains 10 tokens; the first 5 are yellow and labeled '80p', and the last 5 are blue and labeled '80p'. Below the yellow tokens are 5 padlock icons, and below the blue tokens are 5 open lock icons. A hand icon is positioned above the blue tokens, with a horizontal arrow pointing left towards the yellow tokens. Below the hand icon, a horizontal double-headed arrow is shown, with the text 'Player C currently decides how many tokens to protect.' below it.

Player C currently decides how many tokens to protect.

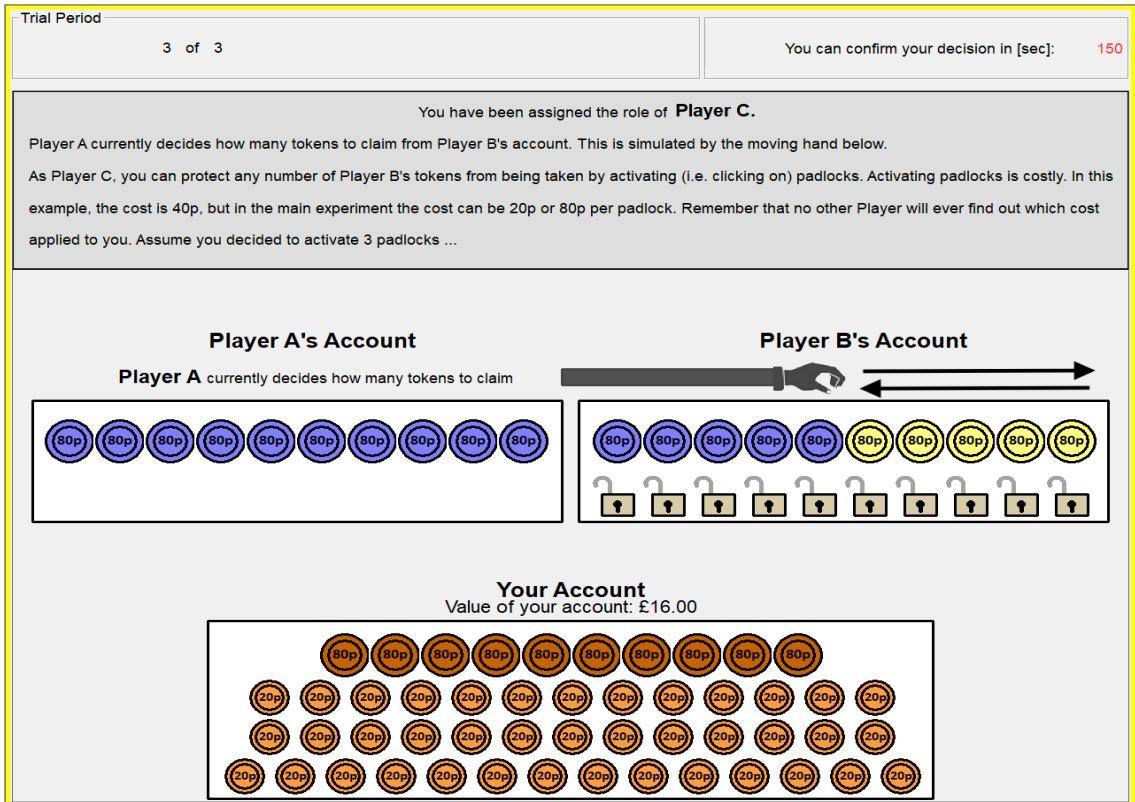
Note: Arrows were not part of the original screen and were added to illustrate motion.

Player C Intro Screen

You will next see Player C's screen

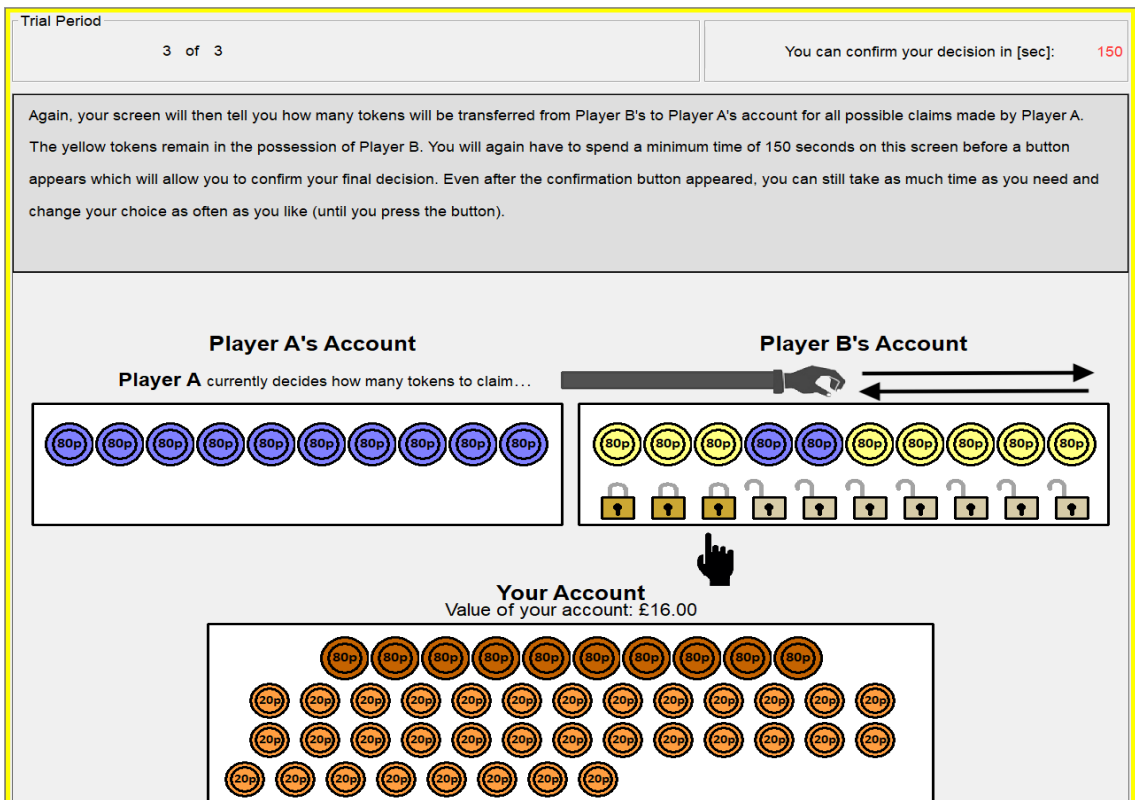
The instructions continue in [sec]: 2

Player C Screen



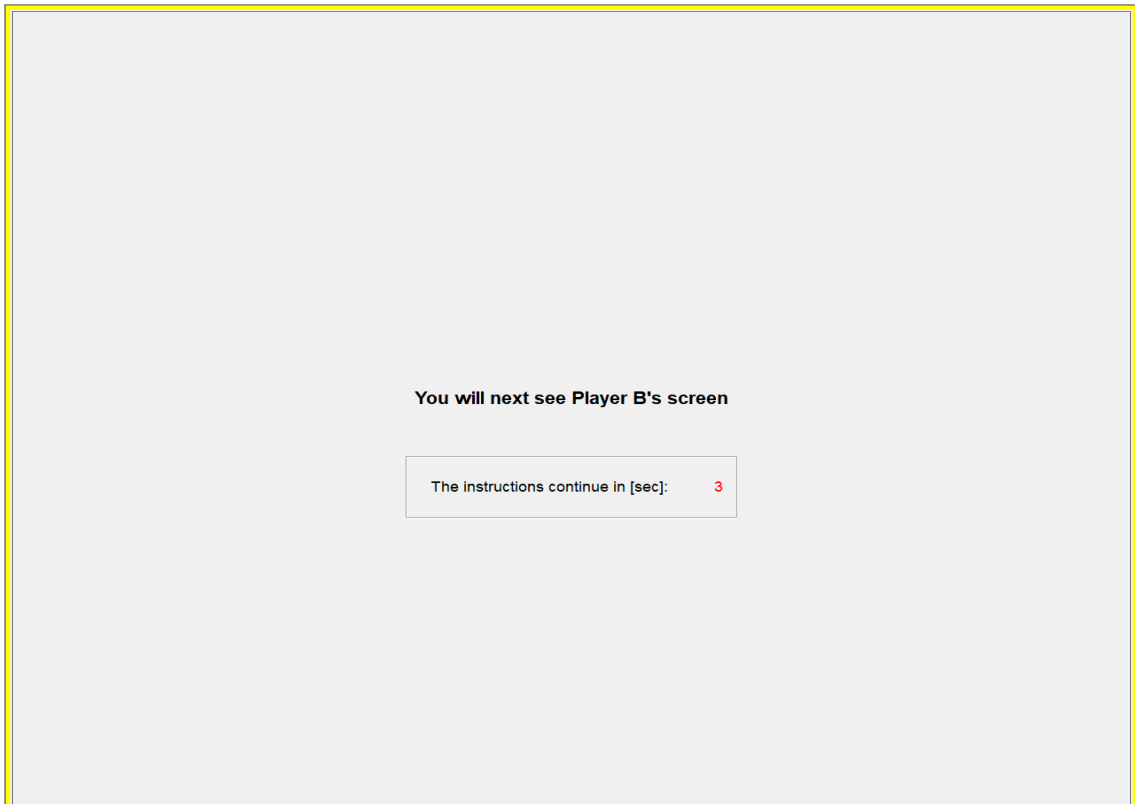
Note: Arrows were not part of the original screen and were added to illustrate motion.

Player C Screen (cont.)

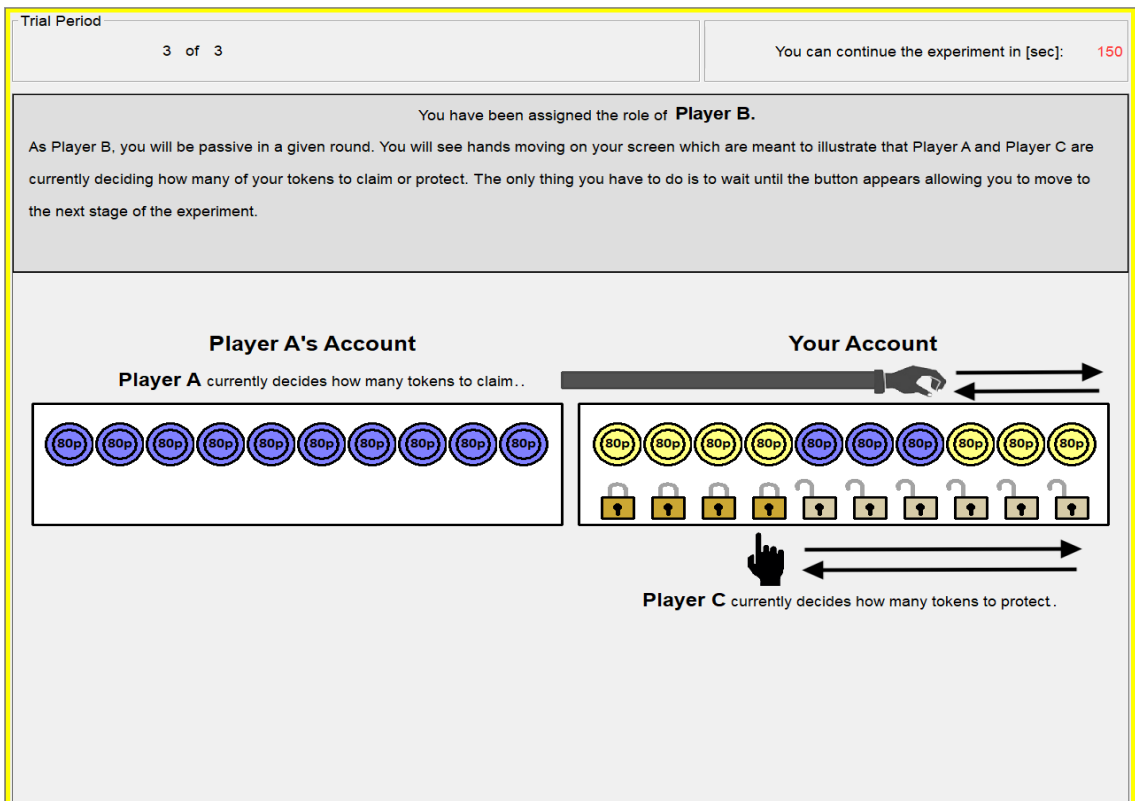


Note: Arrows were not part of the original screen and were added to illustrate motion.

Player B Intro Screen

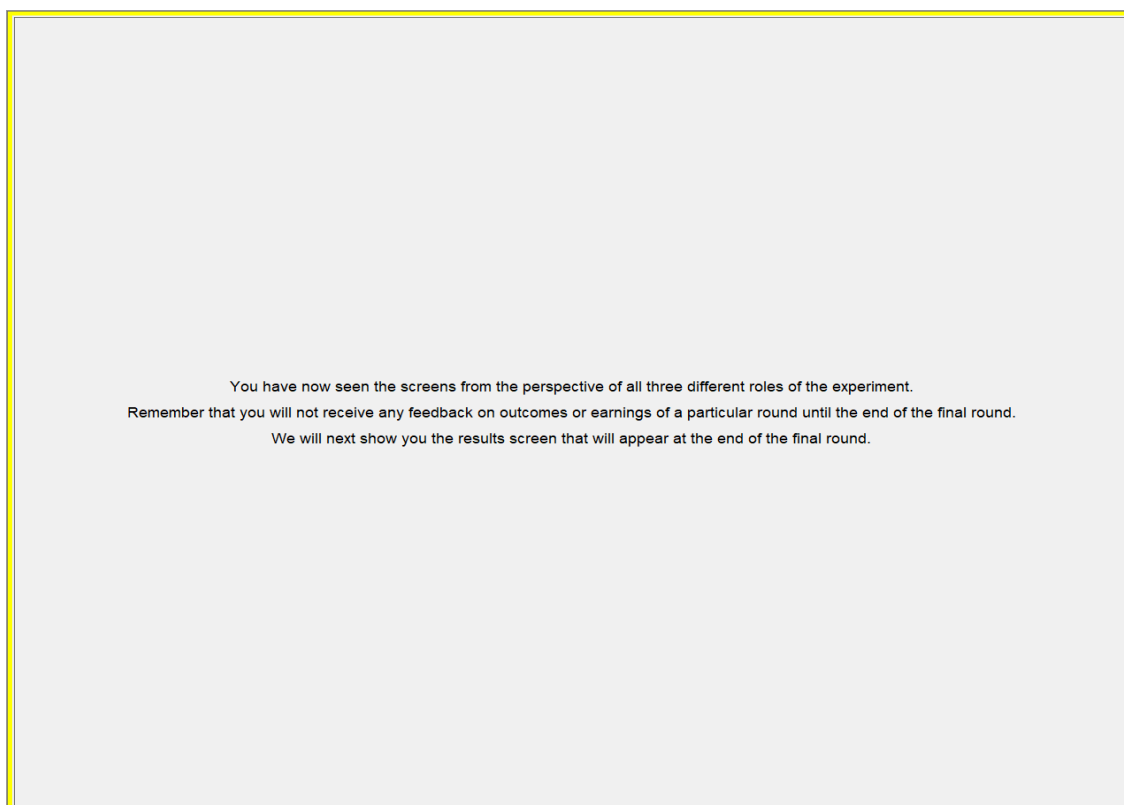


Player B Screen

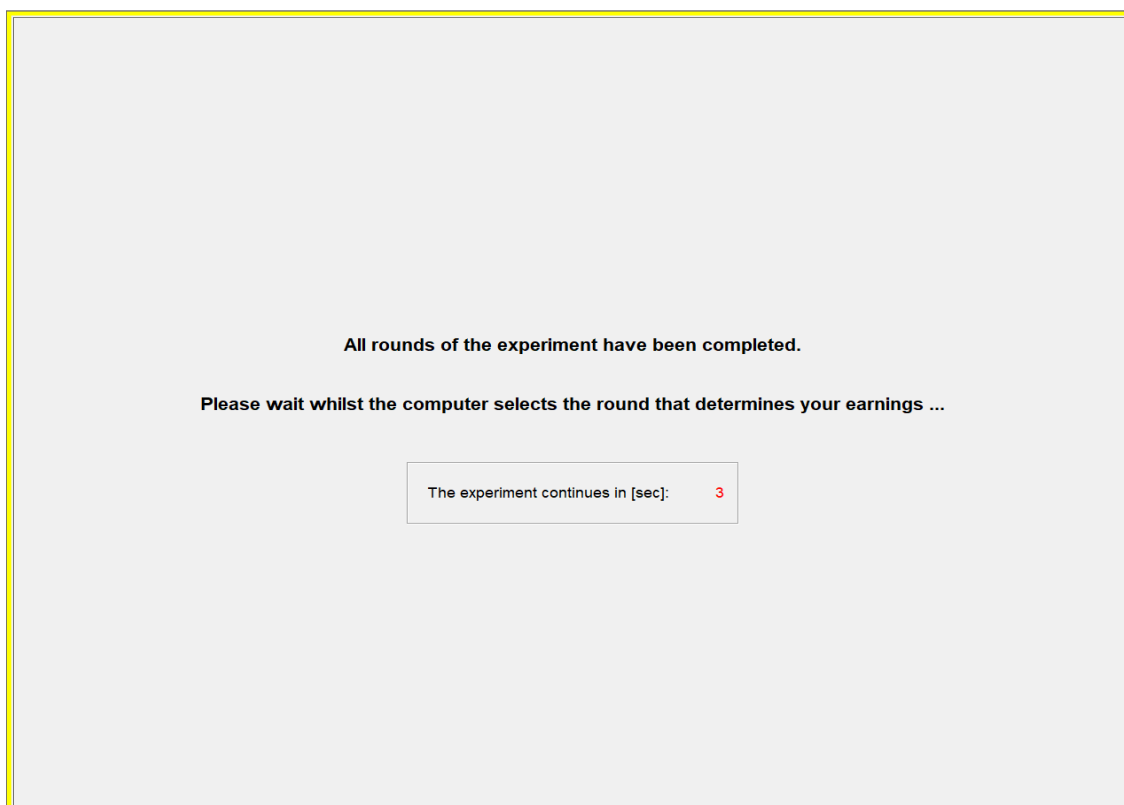


Note: Arrows were not part of the original screen and were added to illustrate motion.

Results Intro Screen



Round Selection Screen – Example 1



Results Screen – Example 1

Results

The computer selected **Round 3** to determine your earnings.

Your role in Round 3 was: **Player A**.

Outcomes in Round 3:
Player A claimed 7 tokens.
Player C protected 3 tokens.
The transfer from B to A is: 4 tokens.
Your earnings: £8 (endowment) + 4 * £0.80 (transfer) = £11.20.

Your Account

Player B's Account

Round Selection Screen – Example 2

All rounds of the experiment have been completed.

Please wait whilst the computer selects the round that determines your earnings ...

The experiment continues in [sec]: 3

Results Screen – Example 2

Results

The computer selected **Round 1** to determine your earnings.

Your role in Round 1 was: **Player B**.

Outcomes in Round 1:

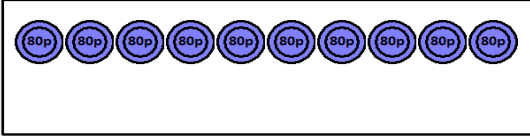
Player A claimed 6 tokens.

Player C protected 6 tokens.

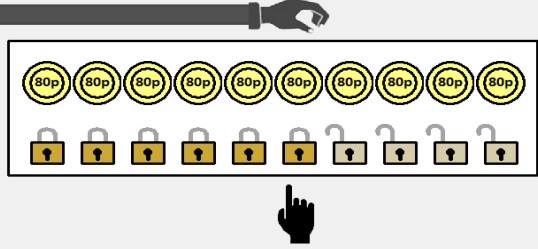
The transfer from B to A is: 0 tokens.

Your earnings: £8 (endowment) - 0 * £0.80 (transfer) = £8.

Player A's Account



Your Account



Round Selection Screen – Example 3

All rounds of the experiment have been completed.

Please wait whilst the computer selects the round that determines your earnings ...

The experiment continues in [sec]: 3

Results Screen – Example 3

Results

The computer selected **Round 2** to determine your earnings.

Your role in Round 2 was: **Player C.**

Outcomes in Round 2:

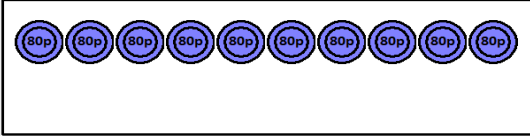
Player A claimed 3 tokens.

Player C protected 6 tokens.


The transfer from B to A is: 0 tokens.

Your earnings: £16 (endowment) - 6 * £0.40 (protection cost) = £13.60.

Player A's Account

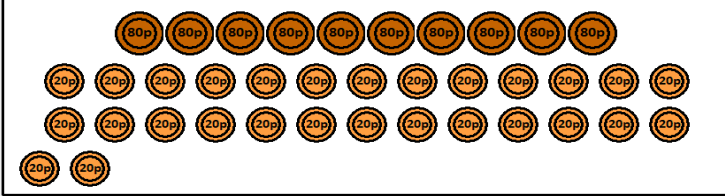


Player B's Account



Your Account

Value of your account: £13.60



2.B.3 Control Questions

Control Question 1:

The data generated in this experiment ...

- ✓ is anonymous, neither the experimenter nor other participants will be able to link my behaviour to me as a person.

links my behaviour in the experiment to me as a person.

links my behaviour in the experiment to me as a person, but only the experimenter will be able to make this connection.

Control Question 2:

The experiment has three rounds. Which of the following is true?

I will play different roles, but will interact with the same participants in a group in all rounds.

- ✓ I will play different roles, and interact with different participants in a group in all rounds.

I will play the same role, but interact with different participants in a group in all rounds.

Control Question 3:

The sequence of decisions in the experiment is as follows ...

First, Player A chooses how many tokens to claim, then Player C decides how many tokens to protect.

First, Player C chooses how many tokens to protect, then Player A decides how many tokens to claim.

- ✓ Player A and Player C decide simultaneously (at the same time) how many tokens to claim and protect.

Control Question 4 (Experiment 1):

Player C's cost of activating padlocks in the experiment can either be 20p or 80p. Which of the following is true?

- ✓ Other players will never be informed which cost (20p or 80p) applied to Player C.

All players will be informed which cost (20p or 80p) applied to Player C.

Only at the end of the experiment will all players be informed which cost (20p or 80p) applied to Player C.

Control Question 4 (Experiment 2):

Player C makes protection decisions under each of 10 possible scenarios. Which of the following is true?

One decision will be implemented randomly.

- ✓ The decision under the true scenario will be implemented.

Player C can choose which decision to implement.

2.B.4 Experiment 2 Instructions

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

In this experiment, a task will be performed for three rounds. At the beginning of the first round, you will randomly be assigned one of three possible roles: Player A, Player B, or Player C. You will then be allocated to a group which includes one Player A, one Player B, and one Player C.

At the beginning of each of the following two rounds, your group and your role in the experiment will change. You will be assigned to an entirely new group of participants in each round which means that you will never be matched with any other participant in the room for more than one round. Moreover, you will be assigned a different role in each of the three rounds. At the end of the experiment, every participant will have played once as Player A, once as Player B, and once as Player C.

In every round, you will start with a freshly generated amount of ‘tokens’ in your private account. Depending on your decisions and/or the decisions of your group members, the amount of tokens in your account can change. Every token has a specific monetary value attached to it.

At the end of the final round, the computer will randomly select *one* of the three rounds of the experiment to determine your final earnings. Your earnings will then be equal to the value of all the tokens which you hold in your private account at the end of the selected round.

The Task

In every round, players start with an endowment of tokens. The total value of a player's initial token endowment is given as follows:

Player A: £8

Player B: £8

Player C: £16

In the task, Player A will have a chance to claim tokens from Player B's account and Player C will have a chance to protect Player B's tokens from being claimed. When we say claim, we mean trying to take. Tokens which are claimed but not protected will be transferred to Player A's account at the end of a given round.

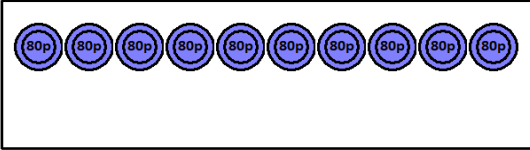
Please have a look on screen where you can see how to claim and protect tokens. You will be told by the experimenter when to return to the paper instructions.

Start of on-screen practice

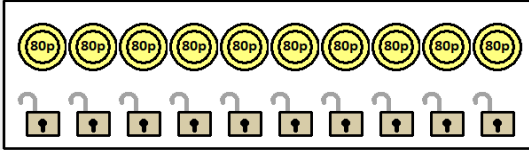
Practicing how to claim tokens (1 of 2):

On the screen, you can see the tokens that Player A and Player B hold in their accounts.
For the moment, please ignore the padlocks that you see on screen, their meaning will be explained soon.
If you are assigned the role of Player A, you will be able to claim tokens from Player B by clicking on them.
Please try to claim 4 tokens by clicking on token number 4 of Player B's account.

Player A's Account



Player B's Account

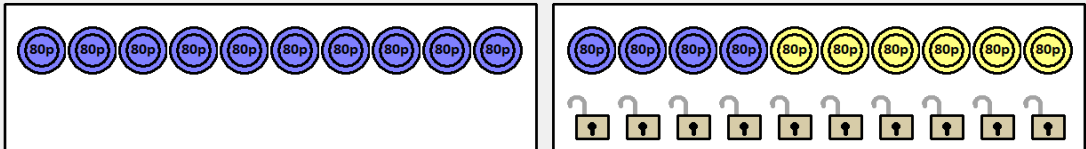


Practicing how to claim tokens (2 of 2):

After you made your claim, you can see that a hand appeared to illustrate your claim.
 All tokens below and to the left of the hand are currently being claimed by you. Claimed tokens change their colour to blue.
 Now, practice a bit by claiming any of the other tokens of Player B. Notice that clicking twice on the same token resets your choice.

Player A's Account

Player B's Account

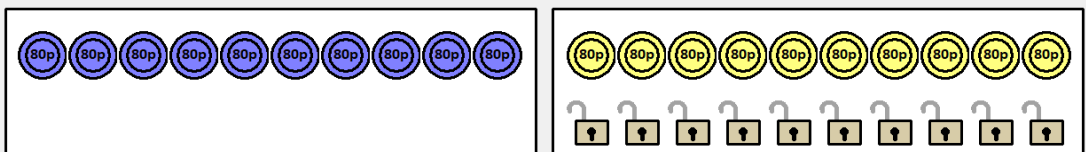


Practicing how to protect tokens (1 of 2):

On the screen, you can see the tokens that Player A, Player B, and Player C hold in their accounts.
 If you are assigned the role of Player C, you will be able to protect Player B's tokens by activating padlocks.
 Activating padlocks is costly for Player C. In this practice stage, the cost is 40p per padlock.
 Please try to activate 6 padlocks by clicking on padlock number 6 of Player B's account.

Player A's Tokens

Player B's Tokens



Your Tokens

Value of your account: £16.00



Practicing how to protect tokens (2 of 2):

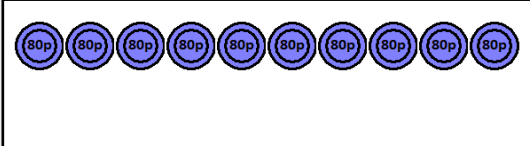
After you clicked on a padlock, you can see that a hand appeared to illustrate your decision.

Activated padlocks click into place and change their colour. All tokens above an activated padlock are currently being protected by you.


Also notice that the cost of activating padlocks (40p) is deducted from your account in real time.

Now, practice a bit by activating any of the other padlocks of Player B. Notice that clicking twice on the same padlock resets your choice.

Player A's Tokens

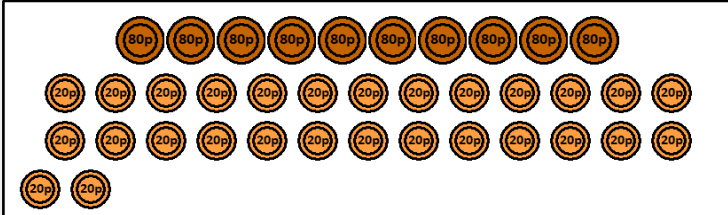


Player B's Tokens



Your Tokens

Value of your account: £13.60



End of on-screen practice

Cost of Protection

You have already seen that activating padlocks is costly for Player C. The cost of activating padlocks will be **80p** per padlock.

Simultaneous Decisions

Player A and Player C will decide *simultaneously*, i.e. at the same time, how many tokens they want to claim or protect. This means that Player A will decide how many of Player B's tokens to claim not knowing how many padlocks Player C will activate. At the same time, Player C will decide how many padlocks to activate not knowing how many tokens Player A will claim.

Protection under Different Scenarios

Suppose you play as Player C in a given round. Your protection decision is implemented in the following way:

At the end of a given round, the computer will calculate the *average* claim made by Players A outside of your own group. Consider the following example: In a given round, six participants played in the role of Player A. Excluding the Player A of your own group, the remaining five A-Players claimed 5, 0, 4, 8, and 7 tokens, respectively. The average claim in this example is 4.8 ($\frac{5+0+4+8+7}{5}$) tokens.

Remember that the average claim is about the behaviour of A-Players *outside* of your group and therefore does not include the Player A from *within* your group. Still, if you knew what the average claim was, this information could be helpful for you in guessing/forecasting what the Player A of your own group would do.

In the decision stage of a given round, you as Player C will be asked to make several protection decisions, one for each of 10 possible scenarios. In every scenario, you will be presented with a different band in which the average claim *could* fall. Note that one (and only one) of the scenarios will be the true scenario, i.e. the scenario that will cover the true average claim.

In a given round and for a given group, only one protection decision will be implemented, namely the one made under the true scenario.

Even though you do not know which scenario will be the true one when deciding, we will ask you to think of each scenario as if it was true. Note that it is sensible for you to treat each scenario as if it was true because amongst the 10 scenarios, one will indeed be true and therefore affect payoffs in a given round.

Determining Payoffs in a Given Round

After all decisions have been made in a given round, the computer calculates the average claims. Next, the computer implements Player C's protection decision made under the true scenario (i.e. the scenario that corresponds to the true average claim).

Payoffs are then determined as follows:

Tokens which are claimed but not protected are transferred from Player B's to Player A's account. Therefore,

- Player A's payoff in a given round is his initial endowment of £8 plus the value of tokens transferred from Player B's to Player A's account.
- Player B's payoff in a given round is his initial endowment of £8 minus the value of tokens transferred from Player B's to Player A's account.
- Player C's payoff in a given round is his initial endowment of £16 minus the cost of all padlocks he activated.

Notice that you will not receive any feedback on outcomes in any of the three rounds of the experiment until the end of the final round. Remember that one of the three rounds of the experiment will be selected to determine your earnings at the end of the experiment and you will only receive feedback on outcomes and your personal earnings of that specific round.

Interface of the Experiment

You already practiced how to claim and protect tokens. We will now take you through the particular screens that you will encounter in the experiment in all three roles to further familiarize you with the interface and processes of the experiment.

Please follow along on screen.

2.B.5 Strategy Method Example Screens

Scenario 1 Intro Screen

Trial Period
3 of 3

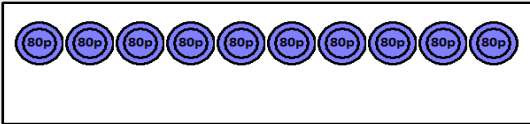
You have been assigned the role of **Player C**.
 Within your own group, Player A is currently deciding how many tokens to claim from Player B's account ... this is simulated by the moving hand below.
 You can protect any number of Player B's tokens from being taken by activating (clicking on) padlocks. The cost per padlock assigned to you is **£0.80**.

You will be asked to provide a separate protection decision for each of 10 possible scenarios.
 Only one of your protection decisions will be implemented, namely the one that corresponds to the true scenario.

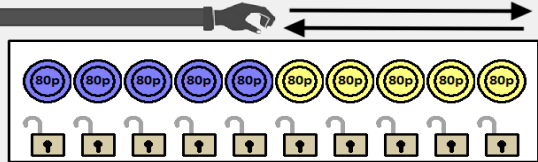
Start Example Scenario

Player A's Account

Player A currently decides how many tokens to claim

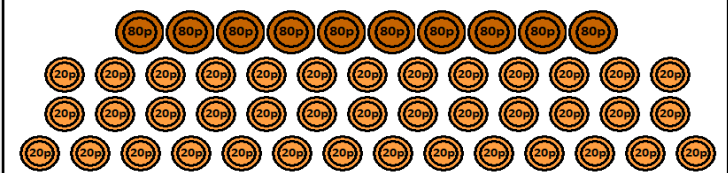


Player B's Account



Your Account

Value of your account: £16.00



Note: Arrows were not part of the original screen and were added to illustrate motion.

Scenario 7 (out of 10) Screen

Trial Period
3 of 3


You can confirm your decision in [sec]: **12**

Scenario 7 (of 10):
 Assume the 5 Players A outside of your own group claimed on average at least 4 but less than 5 (out of 10) tokens. Regarding Player B of your own group, how many of his/her tokens do you want to protect?

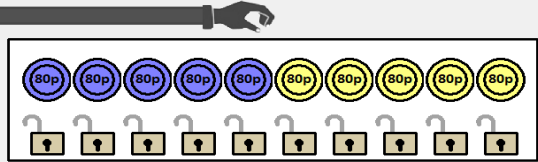
You can protect tokens by activating (clicking on) padlocks. If the assumed average above corresponds to the true average behaviour of Players A, your protection decision made under this scenario will be the one implemented in your group for the current round.
 Assume you decided to activate 3 padlocks under the current scenario ...

Player A's Account

Player A currently decides how many tokens to claim.

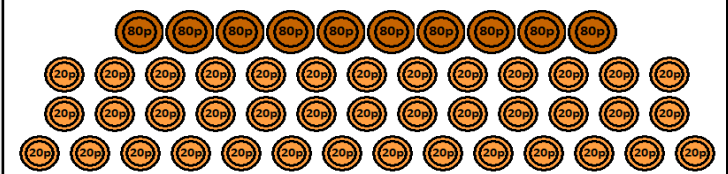


Player B's Account



Your Account

Value of your account: £16.00



Chapter 3:

Costless Sharing, Moral Entitlements and Perception[‡]

[‡]This chapter is based on joint work with Anders Poulsen, Mengjie Wang and Jiwei Zheng. We appreciate financial support from the European Research Council. We would also like to thank the members of the Centre for Behavioural and Experimental Social Science (CBESS) as well as the audience of the 2020 CCC (CBESS-CEDEX-CREED) meeting for their feedback.

3.1 Introduction

Behavioural economic research has documented that people in important economic situations are not motivated exclusively by pecuniary self-interest (see e.g. Camerer, 2003). Many of these findings come from controlled and monetarily incentivised experiments using games such as the Dictator Game (DG) (Forsythe et al., 1994; Engel, 2011). In the typical DG one person, the dictator, is provided with a sum of money, and then decides how much money to give to the other person, the recipient. Average amounts given tend to lie between 20–30% of the surplus (Engel, 2011).

The significant sharing observed in the DG is typically taken as evidence that the dictator cares not only about his own but also the recipient’s money earnings. This has been formalised by models of (outcome based) social preferences (see e.g. Bolton and Ockenfels, 2000; Cooper and Kagel, 2016; Fehr and Schmidt, 1999). Another, not mutually exclusive, interpretation of the data from the DG is that people are what we call *morally motivated*. Their decisions are based on notions such as entitlement, desert, and need (see e.g. Burrows and Loomes, 1994; Gächter and Riedl, 2005; Hoffman and Spitzer, 1985; Rutström and Williams, 2000; Cappelen et al., 2017).

In this paper we measure behaviour and the role of moral reasoning in a new game, the *Costless Sharing Game* (CSG). In the CSG one player, the sharer, produces a valuable resource and decides how much of the resource to *costlessly share* with the recipient. The resource in question is thus non-rival but excludable. In other words, in the CSG the opportunity cost of giving is zero, unlike the DG where it is strictly positive.²⁴ We vary the existence of a moral argument for sharing based on entitlement and desert by varying whether the recipient (a) had to solve the same task that gave rise to the sharer’s resource, or (b) took part as a passive recipient. Following Cappelen et al. (2017), we call this source of motivation *intrinsic moral motivation*.

In addition to this, we seek to understand another factor that may influence the sharer’s costless sharing decision: How much does the recipient know about the sharer’s decision and the context in which it is made? Following Cappelen et al. (2017), we call this source of motivation *extrinsic social motivation*. Previous research (see e.g. Andreoni and Bernheim, 2009; Broberg, Ellingsen and Johannesson, 2007; Dana, Cain and Dawes, 2006; Dana, Weber and Kuang, 2007) has found that knowledge of context and observability of decisions matter since it allows the recipient (and, more generally, a third party or audience) to assess the moral appropriateness of the sharer’s decision. In turn, feelings of pride, guilt, or shame may influence the sharer’s decision making.

²⁴One may also, instead of zero opportunity cost of giving, think of our game as capturing a situation where the costs of giving are very small.

We think the empirical relevance of costless sharing is significant. Examples include emailing presentation slides, sharing documents, and more generally sharing valuable information, knowledge, and advice with someone else.²⁵ While it typically has been costly to produce or obtain these resources in the first place, once they are there it is free to share them with other people. We believe that, surprisingly, we do not currently have any data that allows us to answer the question, *how much will be shared?* Will moral and social arguments remain to be important, or will subjects seek efficiency even in the absence of such arguments, as a result of sharing being costless? Of course, in the real world sharing decisions will depend on a myriad of contextual and institutional factors. These include: social distance, strategic factors based on repeated game interaction, signalling and reputation building, the presence of a principal who can condition monetary rewards and punishment on sharing, and so on. We think all these factors can be studied in future work.

3.2 Related Literature

In this section, we first review closely related studies on intrinsic and extrinsic motivation in the dictator game, where giving is costly. Subsequently, we focus on the existing literature on costless sharing.

3.2.1 Dictator Game Giving

Moral reasoning in the DG has been documented in experiments where the dictator produces the surplus (see e.g. Cherry, Frykblom and Shogren, 2002; Cherry and Shogren, 2008; Oxoby and Spraggon, 2008; Carlsson, He and Martinsson, 2013; Rodriguez-Lara and Moreno-Garrido, 2012; Cappelen et al., 2017; Thunström et al., 2016), instead of receiving it exogenously (windfall income, or “manna from heaven”). For example, Cherry, Frykblom and Shogren (2002) find that the dictator gives substantially less (often zero) to the recipient when the dictator has generated the surplus by performing a real effort task, compared to when the surplus has been provided by the experimenters. Oxoby and Spraggon (2008) find the same, and also show that when the recipient has produced the surplus, the recipient is given much more than in the usual DG (see also Carlsson, He and Martinsson, 2013; Cherry and Shogren, 2008; Ruffle, 1998). Cappelen et al. (2017) observe that the dictator gives more when both have performed a real effort task, compared to when only the dictator has. Cherry and Shogren (2008) as well as Mittone and Ploner (2012) show that not only the legitimisation of assets through effort but also the perceived

²⁵In many cases, sharing presentation slides, documents, advice etc. is beneficial to the sender; in many other cases it is costly. We model an environment where the sharing of resources is beneficial to the recipient and neither beneficial nor costly to the sharer which allows us to investigate whether subjects are averse to sharing per se.

deservingness of receivers play an important role explaining the discussed findings. These studies report behaviour in line with a moral motivation for giving.

In contrast to an intrinsic moral motivation, evidence also suggests that many people are motivated by extrinsic social motivation which is our second area of interest. A dictator who is known by the recipient (or others) to have had the opportunity to give in a situation where giving appears to be morally justified and who did not do so may feel shame or guilt, and anticipating such feelings may lead the dictator to give more, compared to when the recipient would not know that the dictator had an opportunity to give. Giving may also be motivated by pride which results from leaving a positive impression on others. There is now a well-established literature documenting that observability and extrinsic motivation matter (see e.g. Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Ariely, Bracha and Meier, 2009; Bohnet and Frey, 1999; Bursztyn and Jensen, 2017; Chaudhuri, 2011; Dana, Cain and Dawes, 2006; Ekström, 2012; Fehr and Gächter, 2000; Rege, 2004; Rege and Telle, 2004; Soetevent, 2005; Tadelis, 2011).

A shortcoming of the existing literature is that it focuses almost exclusively on studies where giving is costly. In this paper, we wish to assess whether the same forces that motivate giving in the dictator game also play a role when resources can be shared at no cost.

3.2.2 Costless Sharing

The literature on costless sharing is pretty scarce, especially when compared to the significant research that has been dedicated to the analysis of the dictator game. We believe we are the first to experimentally study the role of moral reasoning and extrinsic social motivation in a situation where a resource has been produced via a real effort task and can be shared at no (or negligible) personal cost.

The closest studies of costless sharing we are aware of use the Generosity Game (GG) (see Güth, 2010; Güth, Levati and Ploner, 2012).²⁶ In this game, the dictator chooses the size p of a “pie”, where his own pie amount is fixed at x with $p \geq x$ and $p \leq \bar{p}$ where \bar{p} is the largest feasible pie size. The dictator thus chooses $p \in [x, \bar{p}]$ where it is assumed that $2x < \bar{p}$ (such that it is possible for the dictator to increase the recipient’s payoff above his own). It is experimentally found that a majority of dictators choose the largest pie size thereby favouring efficient and disadvantageous inequality over inefficient equality.²⁷ Follow-up studies using variations of the GG observe similar results. García-Gallego, Georgantzis and Ruiz-Martos (2019) allow

²⁶In what follows, we describe the Dictator Game version; there is also an Ultimatum Game version, where the recipient can reject the proposer’s suggestion. The Envy Game (Casal et al., 2012; Bäker et al., 2015) is a cousin of the Generosity Game.

²⁷According to a type classification, 44.37% of dictators can be regarded as efficiency seeking, 24.72% as inequality averse, and only 3.47% show competitive preferences.

for costless giving *and* taking in the so-called Heaven Dictator Game and find that the dominance of efficiency seeking persists. Bäker et al. (2014) auction off the proposer and responder roles in a GG. They observe that this makes participants care even more for efficiency than for equality compared to when roles are randomly assigned. An exception stems from a three-person generosity experiment by Güth et al. (2010). Here, it is observed that efficiency seeking falls behind equity seeking if general equality is achievable while the opposite is true if inequality is unavoidable.

Our study differs from the discussed literature by introducing moral arguments for and against sharing. The studies using the GG let all money amounts be exogenously given, so moral arguments involving entitlement or desert are absent. Rather, the dictator must decide whether to implement an equal but inefficient distribution, or an unequal but total earnings maximising one. In contrast, we consider the role of moral reasoning based on salient costly effort from producing the surplus. We hypothesise that the salience of effort costs generates moral entitlements that make people unwilling to share the entire surplus with the other player. We moreover think that the significant generosity observed in previous studies could have been affected by subjects' concerns over how their actions are perceived by other players or the experimenter. Our study is particularly suited to minimise such concerns as we conduct our experiment in an online environment where the anonymity of decisions is strengthened. To obtain insights into the relevance of extrinsic social motivation, we vary the information that recipients receive about the context of the sharer's decision and the origin of the shared surplus.

There is a large literature on helping behaviour and organisational citizenship behaviour in teams, companies, and organisations (see LePine, Erez and Johnson, 2002). Our contribution differs from these in several ways. First, many of these studies rely on questionnaire evidence. Second, the only incentivised experiment that we are aware of (see Danilov, Harbring and Irlenbusch, 2019), considers mutual helping in groups, where helping is costly since it takes time away from other activities. We deliberately focus on costless helping and only allow one person in the group to help. Third, as already mentioned above, important considerations such as reciprocity in helping and repeated interaction are deliberately kept out of our study. These can be considered in future work.

3.3 The Experiment

3.3.1 Design

Subjects in our experiment are assigned to one of two possible roles: sharer or recipient. For logistical reasons, sharers were recruited first; they were able to influence earnings of recipients in a later session of the experiment.

Sharers encountered a production stage and a subsequent distribution phase in their experiment. The task in the production stage was borrowed from Cappelen et al. (2017) asking subjects to tick off numbers in a series of tables for up to 8 minutes. The goal was to reach a minimum performance threshold which we deliberately chose such that it would be straightforward for most subjects to complete the task successfully.²⁸ Completing the task (whether successful or not) was described to subjects as a requirement to proceed with the experiment and to receive their participation fee; we did not mention anything about task related rewards at this stage. After having completed the task successfully, sharers were informed that they received a bonus of 100 experimental tokens for their performance. They were also told that each token was worth £0.05 and that the total worth of their tokens would be paid out to them in pounds, together with their participation fee, at the end of the experiment.

In the subsequent distribution phase, sharers were for the first time informed about the costless sharing opportunity and the existence of recipients. Sharers were told that they could “copy and give” any number x of their earned tokens to the other person, where $x \in [0, 100]$. If a copy was made, it had to be given to the other person; it couldn’t be kept. At this point, sharers were also assigned to one of four treatment conditions which varied (i) whether the recipient would also work on the number task and (ii) the information that the recipient would receive about the sharer’s experiment. Table 3.1 summarises our 2x2 factorial treatment design.

- In the *Both Work* conditions, sharers are told that they will be matched with a recipient who will have successfully completed the same task they did but who will not be rewarded for it. In contrast to this, the *Sharer Works* conditions tell subjects that they will be matched with a recipient who has not participated in the production stage at all.
- In the *Full Information* conditions, sharers are told that recipients will be provided with detailed information regarding the origin of the shared amount and the role played by the sharer. In contrast to this, the *No Information* conditions tell sharers that recipients receive the shared amount without any accompanying explanation.

Table 3.1: Factorial Treatment Design

	Full Information	No Information
Both Work	BW_FI	BW_NI
Sharer Works	SW_FI	SW_NI

²⁸Appendix 3.A contains the instructions and decision screens.

After sharers submitted their sharing decision, we asked them to provide an explanation for why they made that choice. We also collected demographic data on age and gender.

3.3.2 Hypotheses

3.3.2.1 Intrinsic Moral Motivation

Equity theory and theories of desert (see for example Adams, 1965; Hoffman and Spitzer, 1985; Güth, 1994; Konow, 2003; Selten, 1978) stipulate that the resource should be shared in proportion to inputs. In Sharer Works, only the sharer has worked on the number task which means that the recipient can be considered *undeserving* of the surplus. This gives sharers a justification not to share. In Both Work, the recipient can be seen as having exerted the same effort in the number task as the sharer. This, coupled with the fact that the recipient was not rewarded for his effort, makes him *deserving* of being shared with, thus more should be shared. Since these moral arguments do not depend on the information condition (FI or NI), we can state the following hypothesis:

Hypothesis 1: *More will be shared in the BW compared to the SW conditions.*

3.3.2.2 Extrinsic Social Motivation

The findings on moral wiggle room in dictator games (see for example Dana, Weber and Kuang, 2007) show that many dictators give significant amounts, not because they want to (intrinsic moral motivation), but because they feel compelled to (extrinsic social motivation). In the No Information conditions, extrinsic motivation based on shame and pride is ruled out since the recipient receives *no information* about the origin of the received amount or the sharer's experiment. In the Full Information conditions, the receiver does receive *detailed information* about the sharer's role in determining the received amount, thus shame and pride can motivate sharing. We can state the following hypothesis:

Hypothesis 2: *More will be shared in the FI compared to the NI conditions.*

3.3.3 Procedures

The experiment was programmed using oTree (Chen, Schonger and Wickens, 2016) and deployed online using the recruitment platform Prolific (www.prolific.co). We pre-registered the experiment on AsPredicted (www.aspredicted.org) with reference #45868. Participants were current UK residents with English as their first language. The first session was run in August 2020 and involved the recruitment of all 240

sharers. The second session which involved the recipients was run in September 2020. The experiment was combined with an unrelated survey experiment which took place after the CSG in the sharers experiment and before the CSG in the recipients experiment. The CSG in the sharers experiment lasted approximately 10 minutes, average earnings were £6.85, including a participation fee of £2. 53% of our participants were female, the average age was 34 years.

3.4 Results

We start this section by reporting the results of our full sample. We then conduct a supplementary analysis considering only subjects who – based on their control question responses – demonstrated that they understood the instructions of the experiment. Unless otherwise stated, reported Z statistics reflect Wilcoxon rank sum tests (see Siegel and Castellan, 1988).

3.4.1 Main Results

Almost all subjects (97%) successfully completed the numbers task. The average completion time was just under 4 minutes (237 seconds).

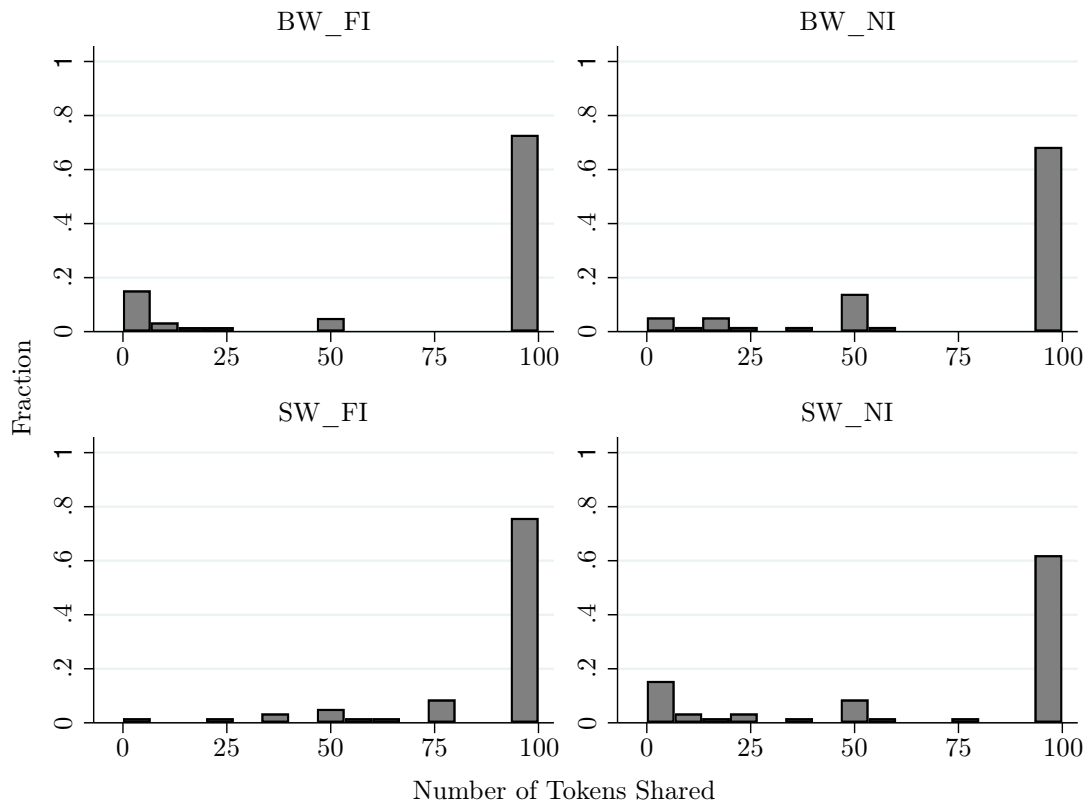
Table 3.2 provides summary statistics of the number of tokens shared in each treatment. The associated distributions are shown in Figure 3.1. Overall, there appears to be little reluctance to share. Our test of hypothesis 1 yields a null result: there are no statistically significant differences in observed sharing between the BW and SW conditions. This holds for the comparison of treatments BW_FI with SW_FI where the observed differences go in the opposite direction of what was predicted (76.6 vs. 89.0; $Z = 0.421$; $p = 0.210$, one-tailed) and for the comparison of treatments BW_NI with SW_NI where the direction is in line with the prediction (78.8 vs. 70.9; $Z = 1.102$; $p = 0.135$, one-tailed). We state the following result:

Result 1. *There is no significant difference in sharing detected between the BW and SW conditions.*

Table 3.2: Summary Statistics of Tokens Shared

Treatment	Mean	Std. Dev.	n
BW_FI	76.6	39.9	59
SW_FI	89.0	22.5	58
BW_NI	78.8	33.6	57
SW_NI	70.9	40.5	58
Total	78.8	35.3	232

Figure 3.1: Distributions of Tokens Shared



With respect to hypothesis 2, we find support in the Sharer Works conditions, not however in the Both Work conditions. There are no significant differences detected between treatments BW_FI and BW_NI (76.6 vs. 78.8; $Z = 0.034$; $p = 0.487$; one-tailed). In contrast, sharing in treatment SW_FI is significantly higher compared to treatment SW_NI (89.0 vs. 70.9; $Z = 2.172$; $p = 0.015$; one-tailed). We state the following result:

Result 2. *Significantly more is shared in treatment SW_FI than SW_NI. There is no significant difference in sharing detected between treatments BW_FI and BW_NI.*

The relatively low sharing observed in treatment BW_FI is somewhat surprising to us, given that the combination of a moral argument for sharing and the provision of full information was expected to generate most sharing. To better understand the motives behind sharing, we investigated how subjects justified their decisions in the written explanations they provided.

Table 3.3 reports the results of a text categorisation. Many text justifications refer to sharing being costless, morally appropriate or kind. As intended by design, moral arguments for sharing based on entitlement are much more pronounced in the BW compared to the SW conditions according to Fisher's exact tests (full information: 28.8% vs. 8.6%, $p < 0.01$; no information: 26.3% vs. 3.4%, $p < 0.01$). As an illustration, subject #85 in the BW condition shares 100 tokens and

Table 3.3: Classification of Written Explanations

	BW_FI	BW_NI	SW_FI	SW_NI
Moral Argument	28.8%	26.3%	8.6%	3.4%
Costless	45.8%	47.4%	60.3%	51.7%
Generous/Kind	52.5%	50.9%	74.1%	67.2%
Reciprocal	10.2%	5.3%	6.9%	1.7%
Surprise	5.1%	3.5%	10.3%	12.1%
Other/Unspecific	6.8%	7.0%	10.3%	12.1%
Misunderstanding	10.2%	10.5%	3.4%	6.9%
n	59	58	57	58

writes “We both did the experiment correctly, but the other person received no tokens so I have equalled us out, hopefully!”. In contrast, subject #58 in the SW condition shares 25 tokens and writes “Because I did something to actually earn mine and they did not. I was still willing to be charitable but not completely”. Since moral arguments for sharing are nearly absent in the SW condition, subjects more often justify their sharing in that condition as an act of generosity or kindness (full information: 74.1% vs. 52.5%, $p = 0.021$; no information: 67.2% vs. 50.9%, $p = 0.089$).²⁹ We don’t find any further between-treatment differences across the remaining categories.

Another insight from the text analysis is that roughly 7.8% (18 of 232) of texts reveal some misunderstanding of the features of the game, most notably related to sharing being costless. A closer look at the sharing behaviour of subjects whose justifications were classified to reveal misunderstanding shows substantially lower sharing amounts compared to the remaining subgroup (26.8 compared to 83.2 shared tokens on average). To obtain a cleaner picture of subjects’ willingness to share which attempts to correct for distortions in the data due to misunderstanding, we next conduct a supplementary analysis focusing on subjects who demonstrated in the control question stage that they understood the rules of the experiment.

3.4.2 Supplementary Results

Before subjects were given the opportunity to share, we presented them with a set of three control questions which tested their understanding of the core features of our design: (i) sharing being costless, (ii) the recipient exerting effort or not, and (iii) the recipient receiving information or not. For the specifics such as the exact phrasing of the questions, we refer the reader to Appendix 3.A.

²⁹ Two out of 115 subjects in the SW conditions perceive the receiver as deserving because he or she was not given an opportunity to work on the task.

Table 3.4: Control Question Attempts

First attempt?	Question 1	Question 2	Question 3
True	77.6%	76.3%	83.2%
False	22.4%	23.7%	16.8%
n	232	232	232

For each control question, we recorded a binary variable in the experiment telling us whether a subject found the correct solution to a given question on the first try or not. Table 3.4 summarises the findings. For the remainder of the analysis, we consider only those subjects who correctly answered all three control questions on the first try, i.e. without making any mistakes. This approach could be considered extreme as it likely overstates the degree of misunderstanding in our experiment by also excluding subjects who learned from their mistakes. We however prefer this approach over excluding subjects based on a subjective evaluation of their writing.

Table 3.5 and Figure 3.2 reproduce the statistics from the previous analysis for the reduced sample. It is evident that for the subsample, there is even less reluctance to share than for the full sample. Again, we observe no statistically significant differences in observed sharing between treatments BW_FI and SW_FI (95.2 vs. 93.1; $Z = 1.235$; $p = 0.158$; one-tailed). However we do find mild evidence that less is shared in treatment SW_NI than BW_NI (90.8 vs. 79.3; $Z = 1.469$; $p = 0.068$; one-tailed).

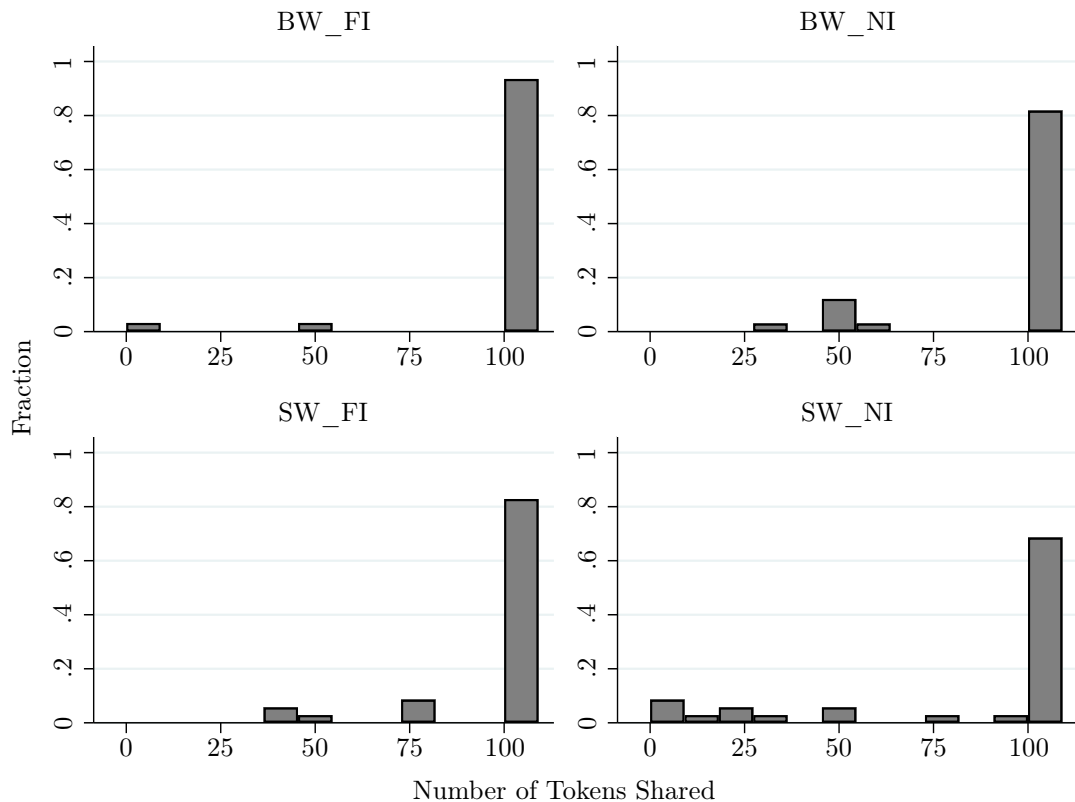
Result 3. (*Reduced Sample*) *There is mild evidence that more is shared in treatment BW_NI than SW_NI. There is no significant difference in sharing detected between treatments BW_FI and SW_FI.*

Regarding the role of information, there is no significant difference induced by providing information in the BW conditions (95.2 vs. 90.8; $Z = 1.334$; $p = 0.145$,

Table 3.5: Summary Statistics of Tokens Shared (Reduced Sample)

Treatment	Mean	Std. Dev.	n
BW_FI	95.2	19.1	31
SW_FI	93.1	17.0	35
BW_NI	90.8	20.2	33
SW_NI	79.3	35.9	35
Total	90.4	25.1	134

Figure 3.2: Distributions of Tokens Shared (Reduced Sample)



one-tailed). There however is a mild difference when considering the SW conditions (93.1 vs. 79.3; $Z = 1.624$; $p = 0.049$, one-tailed).

Result 4. (*Reduced Sample*) *There is mild evidence that more is shared in treatment SW_FI than SW_NI. There is no significant difference in sharing detected between treatments BW_FI and BW_NI.*

The results of the supplementary analysis have to be considered with caution as they are based on fewer observations. Nevertheless, the findings are by and large consistent with those obtained when considering the full sample. First of all, there is very little reluctance to share when sharing is costless. Second of all, there is evidence that the combination of the SW and NI features of our experiment reduces sharing, albeit by a small amount.

3.5 Discussion

The literature on costless sharing is pretty scarce and has predominantly considered environments where moral arguments for sharing played a negligible role. In the closely related Generosity Game for example, subjects are endowed with “mana from heaven”; it therefore comes as no surprise to us that subjects in the GG tend to be remarkably generous and show very little reluctance to increase a recipient’s

earnings. In our experiment, sharers had to earn their endowments with their effort, thereby strengthening the legitimacy of their assets. In addition to this, we varied the deservingness of the recipient by setting up conditions where the recipient had to exert a similar effort to that of the sharer as opposed to having exerted no effort at all. Despite our deservingness manipulations, we observe very little reluctance to costlessly share earned wealth with a recipient. Quite interestingly however, we do observe a tendency for sharing to be lower in treatments where the receiver is *both* undeserving and receives no information about the existence of the sharer. A similar interaction effect has been observed by Cappelen et al. (2017) in the context of a dictator game where sharing was costly. In their experiment, extrinsic social motivation plays out more strongly when the dictator perceives that there exists a moral argument for giving and we think the same may be true in our costless sharing experiment.

We think follow-up research could investigate whether a stronger reluctance to costlessly share could be observed under conditions which push the perception of receiver undeservingness even further, e.g. by implementing a protocol similar to that of Cherry and Shogren (2008) who let the recipient be someone who actively decided to opt out of the experiment. Another exciting research project could apply techniques similar to those used in the previous chapters of this thesis to investigate whether opportunities to self-deceive about the cause of a missed sharing opportunity decrease sharing.

3.6 Conclusion

We presented the results of an online experiment which implemented the Costless Sharing Game. In this game, a sharer first earns a resource by completing an effort task and is then offered the opportunity to share the resource at no personal cost with another person, the recipient. We used the CSG to consider how the amount shared depends on moral reasoning based on entitlement and desert (“intrinsic moral motivation”) and on whether the context of the decision of the sharer is known by the recipient (“extrinsic social motivation”).

Our results indicate that the remarkably high generosity observed in previous experiments which allowed subjects to increase others’ earnings at no personal cost extends to environments such as ours, where moral arguments for sharing are manipulated. Interestingly, we also found mild evidence of an interaction between our treatment conditions which indicates less sharing when neither intrinsic moral nor extrinsic social arguments for sharing are present.

We think that especially the latter finding warrants further research as to whether such an interaction is a robust feature of costless sharing. But so far, the evidence points towards remarkably little reluctance to share when sharing costs are removed.

Appendix for Chapter 3

3.A Instructions and Screens

Consent Screen

Information about the study

If you volunteer to do so, you will participate in a study about attitudes and behaviour that will last approximately 30 minutes. You will not be subjected to any risks other than those normally associated with using Prolific.

If you successfully complete the study, you will receive a participation fee of £4.

As part of the study, you will be asked to provide some basic demographic information. You will not be asked to provide any sensitive personal data. The data will be used only for research purposes and may be reported in research publications and made available to other researchers. These data will be entirely anonymous.

Participation in the study is voluntary. You have the right to stop at any time without giving any reason.

This study has been approved by the Research Ethics Committee of the School of Economics at the University of East Anglia in accordance with the policies of the European Research Council. The data will be handled in line with current data protection legislation.

Consent to participate

I have read and understood the 'Information about the study' text and agree to participate in the study.

[Proceed](#)

Instructions Screen

Instructions

Welcome to this experiment and thank you for participating.

This experiment consists of two parts: a *number recognition task* and a *survey*. The two parts are completely unrelated.

Each part will last approximately 15 minutes and you will receive a participation fee of £4 for completing *both* parts of the experiment. You must complete both parts in order to receive any payment. You may be excluded from any payments if it becomes evident that you have not followed the instructions of the experiment.

Your identity in this experiment is anonymous. This means it will not be possible for other participants or the researchers conducting the experiment to relate the decisions you make in the experiment to you as a person.

If you have queries regarding the experiment, you can send an email to Kevin Grubiak (k.grubiak@uea.ac.uk) who is administratively responsible for the experiment.

Please enter your prolific ID here:

[Start Part 1](#)

Number Task Example

Number Recognition Task - Example

In this part of the experiment you will be given 8 minutes to complete a task. The task is to find certain numbers in tables that have many different numbers. An example table is shown below. In the actual task you will face larger tables.

You get one point each time you tick off the correct number in the table and you lose one point each time you tick off an incorrect number. You can move to a new table at any time by clicking 'Next Table'. To complete the task, you will have to collect 40 points within the time limit of 8 minutes. You will be informed about the number of points you have earned during the task. The task will end when you reach the target or when you are out of time. You will then receive further instructions.

To familiarise yourself with the task, please tick off all numbers highlighted in red. Notice that the number to look for is **384** and by ticking off the highlighted numbers you are ticking off 3 correct numbers (384) and 1 incorrect number (143). Click 'Next Table' to submit your response for the current table.

400	582	365	618	285	949	339	384	429	339	229	235
142	384	865	195	497	632	242	585	591	197	311	302
841	384	261	389	143	292	121	708	339	368	384	807
346	697	826	765	632	447	750	384	384	798	948	191
296	903	856	665	384	704	348	989	959	106	842	229
940	193	685	223	304	558	292	684	228	768	389	854

Number to be ticked off: 384

Your progress: 0/40 points

Next Table

Number Task Example (cont.)

Number Recognition Task - Example

As you can see in the illustration below, you earned 2 points. This is because you ticked off 3 correct and 1 incorrect numbers in the previous table. You can also see below that a new table and a new number to search for was generated. If this was the real task, you would now continue with the task by ticking off the new number (719).

By clicking the button below, you start the real task. You will have 8 minutes to reach the target.

Start Task

371	356	719	768	807	719	719	776	215	864	217	871
934	481	579	158	719	605	719	666	185	462	873	462
903	940	169	716	358	266	678	362	635	993	993	576
825	719	852	719	298	248	579	295	370	276	185	710
736	639	817	665	768	100	297	719	402	668	467	719
857	919	765	102	646	783	719	478	912	819	222	130

Number to be ticked off: 719

Your progress: 2/40 points

Next Table

Number Task

Number Recognition Task

Time left to complete the task: **4:33**

You earn one point for each time you tick off a correct number, and you lose one point each time you tick off an incorrect number.

Your progress: 14/40 points

Number to be ticked off: 592

570	573	665	573	771	771	573	937	769	592	579	579	771	579	570	771	943	771	592	937	769	580	771	683	665	943	139	139
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
769	943	592	661	592	683	139	592	937	771	771	661	665	580	580	570	573	943	570	943	592	573	937	771	661	943	579	771
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
570	389	592	580	937	579	592	580	573	579	937	665	771	769	661	592	570	282	937	389	661	771	139	579	580	937	389	570
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
570	665	592	592	139	937	769	592	769	580	282	661	389	661	579	661	771	282	769	389	683	139	937	570	937	683	937	937
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
282	389	665	665	282	282	592	580	771	661	937	665	579	580	282	665	580	580	683	937	570	943	580	579	769	937	661	661
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
771	661	389	579	592	580	139	769	592	937	769	661	580	570	570	665	570	592	661	665	769	282	683	579	592	282	937	573
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
937	661	937	282	580	771	943	937	771	592	579	769	592	389	771	661	139	389	389	573	937	570	683	769	592	665	943	769
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Next Table

Number Task Success

Results

You have *successfully* completed the number recognition task and for this we are awarding you a **bonus of 100 tokens**.

Each token has a value of 5 pence, so your 100 tokens are worth 5 pounds. At the end of the experiment, your 100 tokens will be converted into 5 pounds which will be paid to you, together with your 4 pounds participation fee.

Proceed

Sharing Introduction (BW_FI)

Opportunity to share

You will now be given the opportunity to make copies of your tokens, and to give them to another person who will take part in a future session of this experiment. Just like the current experiment, the other person's experiment will consist of the number recognition task and the survey which that person agreed to complete in exchange for a £4 participation fee. The person that you will be matched with will be selected randomly amongst all participants who - just like you - successfully completed the number recognition task. In contrast to you, this other person will not be awarded any bonus for having completed the number recognition task successfully.

It does not cost you anything to give copies of your tokens to the other person, and you keep the tokens you earned. So you keep all your 100 tokens regardless of how many you copy. If you make a copy of a token, you must give it to the other person; you cannot keep the copy yourself.

After the other person completes their own experiment, he or she will receive 5 pence for every token you decided to copy and give. The other person will also receive detailed information about your experiment and the choices that you made. The exact message that this person will receive is the following:

'In addition to your participation fee for today's experiment, you are receiving £x. This amount was determined by a participant in a previous session of this experiment. This participant encountered the same stages you did in the experiment. In contrast to you, this participant received a bonus of £5 for having successfully completed the number recognition task. The bonus was expressed in 100 tokens (each worth 5 pence). The participant was then offered to copy and give you any number of tokens in the range of 0 to 100. It did not cost the participant anything to copy or give you tokens, but once copied these tokens had to be given to you; they couldn't be kept. Before the participant made a decision, he or she was shown a copy of this message. The participant decided to give you x token copies.'

Before you make your decision, we ask you to answer a few questions to ensure you understood the information provided above.

[Start Questionnaire](#)

Sharing Introduction (SW_FI)

Opportunity to share

You will now be given the opportunity to make copies of your tokens, and to give them to another person who will take part in a future session of this experiment. In contrast to the current experiment, the other person's experiment will only consist of the survey part which that person agreed to complete in exchange for a £2 participation fee. The person that you will be matched with will be selected randomly amongst all participants who will take part in the future experiment. In contrast to you, this other person will not work on the number recognition task and will therefore not be awarded any bonus.

It does not cost you anything to give copies of your tokens to the other person, and you keep the tokens you earned. So you keep all your 100 tokens regardless of how many you copy. If you make a copy of a token, you must give it to the other person; you cannot keep the copy yourself.

After the other person completes their own experiment, he or she will receive 5 pence for every token you decided to copy and give. The other person will also receive detailed information about your experiment and the choices that you made. The exact message that this person will receive is the following:

'In addition to your participation fee for today's experiment, you are receiving £x. This amount was determined by a participant in a previous session of this experiment. This participant encountered the same survey you did but also had to solve additional tasks. In one of these tasks, this participant received a bonus of £5 for having successfully completed a number recognition task. The bonus was expressed in 100 tokens (each worth 5 pence). The participant was then offered to copy and give you any number of tokens in the range of 0 to 100. It did not cost the participant anything to copy or give you tokens, but once copied these tokens had to be given to you; they couldn't be kept. Before the participant made a decision, he or she was shown a copy of this message. The participant decided to give you x token copies.'

Before you make your decision, we ask you to answer a few questions to ensure you understood the information provided above.

[Start Questionnaire](#)

Sharing Introduction (BW_NI)

Opportunity to share

You will now be given the opportunity to make copies of your tokens, and to give them to another person who will take part in a future session of this experiment. Just like the current experiment, the other person's experiment will consist of the number recognition task and the survey which that person agreed to complete in exchange for a £4 participation fee. The person that you will be matched with will be selected randomly amongst all participants who - just like you - *successfully completed* the number recognition task. In contrast to you, this other person will not be awarded any bonus for having completed the number recognition task successfully.

It does not cost you anything to give copies of your tokens to the other person, and you keep the tokens you earned. So you keep all your 100 tokens regardless of how many you copy. If you make a copy of a token, you must give it to the other person; you cannot keep the copy yourself.

After the other person completes their own experiment, he or she will receive 5 pence for every token you decided to copy and give. The other person will receive no information about your experiment or the choices that you made. The exact message that this person will receive is the following:

'In addition to your participation fee for today's experiment, you are receiving £x.'

If you decide to share 0 tokens, the other participant will not receive any message.

Before you make your decision, we ask you to answer a few questions to ensure you understood the information provided above.

[Start Questionnaire](#)

Sharing Introduction (SW_NI)

Opportunity to share

You will now be given the opportunity to make copies of your tokens, and to give them to another person who will take part in a future session of this experiment. In contrast to the current experiment, the other person's experiment will only consist of the survey part which that person agreed to complete in exchange for a £2 participation fee. The person that you will be matched with will be selected randomly amongst all participants who will take part in the future experiment. In contrast to you, this other person will not work on the number recognition task and will therefore not be awarded any bonus.

It does not cost you anything to give copies of your tokens to the other person, and you keep the tokens you earned. So you keep all your 100 tokens regardless of how many you copy. If you make a copy of a token, you must give it to the other person; you cannot keep the copy yourself.

After the other person completes their own experiment, he or she will receive 5 pence for every token you decided to copy and give. The other person will receive no information about your experiment or the choices that you made. The exact message that this person will receive is the following:

'In addition to your participation fee for today's experiment, you are receiving £x.'

If you decide to share 0 tokens, the other participant will not receive any message.

Before you make your decision, we ask you to answer a few questions to ensure you understood the information provided above.

[Start Questionnaire](#)

Control Questions

Questionnaire

Q1: You earned 100 tokens. Suppose you decided to make x token copies. Which of the following is true?

- You keep the copied tokens and therefore have $100 + x$ tokens.
- You give x token copies to the other person and are left with $100 - x$ tokens.
- You keep your 100 tokens and give x token copies to the other person.

Q2: The person that you are matched with:

- Will have completed the number recognition task successfully and receive a bonus of 100 tokens for it.
- Will have completed the number recognition task successfully and receive no bonus for it.
- Will not have worked on the number recognition task and therefore not receive a bonus.

Q3: The person that you are matched with:

- Will receive no information regarding your experiment or the origin of any tokens that this person receives from you.
- Will receive detailed information regarding your experiment and the origin of any tokens that this person receives from you.

[Submit Questionnaire](#)

Sharing Decision

Sharing decision

Please decide how many of your 100 tokens you want to copy and give to the other person (any number between and including 0 and 100 is allowed):

I decide to copy and give token(s) to the other person.

[Submit](#)

Sharing Explanation

Explanation

You decided to copy and give 36 tokens to the other person.

Please explain why you made this decision:

Type here.

Submit

Earnings Screen

You completed the experiment

You earned a participation fee of 4 pounds and a bonus of 5 pounds.

Proceed

Bibliography

- Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman.** 2011. "Reference Points and Effort Provision." *American Economic Review*, 101(2): 470–492.
- Adams, J. Stacy.** 1965. "Inequity in Social Exchange." *Advances in Experimental Social Psychology*, 2: 267–299.
- Akerlof, George A., and Rachel E. Kranton.** 2000. "Economics and Identity." *The Quarterly Journal of Economics*, 115(3): 715–753.
- Akerlof, George A., and Rachel E. Kranton.** 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton University Press.
- Andreoni, James, and B. Douglas Bernheim.** 2009. "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica*, 77(5): 1607–1636.
- Andreoni, James, and Ragan Petrie.** 2004. "Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising." *Journal of Public Economics*, 88(7): 1605–1623.
- Ariely, Dan, Anat Bracha, and Stephan Meier.** 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review*, 99(1): 544–555.
- Bäker, Agnes, Werner Güth, Kerstin Pull, and Manfred Stadler.** 2014. "Entitlement and the Efficiency-equality Trade-off: an Experimental Study." *Theory and Decision*, 76(2): 225–240.
- Bäker, Agnes, Werner Güth, Kerstin Pull, and Manfred Stadler.** 2015. "The Willingness to Pay for Partial vs. Universal Equality: Insights from Three-person Envy Games." *Journal of Behavioral and Experimental Economics*, 56: 55–61.
- Bardsley, Nicholas.** 2008. "Dictator Game Giving: Altruism or Artefact?" *Experimental Economics*, 11(2): 122–133.

- Baumeister, Roy F., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice.** 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology*, 74(5): 1252–1265.
- Bem, Daryl J.** 1972. "Self-Perception Theory." In *Advances in Experimental Social Psychology*. Vol. 6, 1–62. Elsevier.
- Bénabou, Roland, and Jean Tirole.** 2004. "Willpower and Personal Rules." *Journal of Political Economy*, 112(4): 848–886.
- Bénabou, Roland, and Jean Tirole.** 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652–1678.
- Bénabou, Roland, and Jean Tirole.** 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics*, 126(2): 805–855.
- Bhattacharya, Puja, and Arjun Sengupta.** 2016. "Promises and Guilt." Available at SSRN: <https://ssrn.com/abstract=2904957>.
- Bicchieri, Cristina.** 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.
- Bicchieri, Cristina, and Eugen Dimant.** 2019. "Nudging with Care: The Risks and Benefits of Social Information." *Public Choice*, 1–22.
- Bicchieri, Cristina, Eugen Dimant, and Silvia Sonderegger.** 2020. "It's Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs." Available at SSRN: <https://ssrn.com/abstract=3326146>.
- Bodner, Ronit, and Drazen Prelec.** 2003. "Self-Signaling and Diagnostic Utility in Everyday Decision Making." *The Psychology of Economic Decisions*, 1: 105–26.
- Bohnet, Iris, and Bruno S. Frey.** 1999. "The Sound of Silence in Prisoner's Dilemma and Dictator Games." *Journal of Economic Behavior & Organization*, 38(1): 43–57.
- Bolton, Gary E., and Axel Ockenfels.** 2000. "ERC: A Theory of Equity, Reciprocity and Competition." *The American Economic Review*, 90(1): 166–193.
- Brandts, Jordi, and Gary Charness.** 2011. "The Strategy versus the Direct-Response Method: a First Survey of Experimental Comparisons." *Experimental Economics*, 14(3): 375–398.

-
- Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson.** 2007. "Is Generosity Involuntary?" *Economics Letters*, 94(1): 32–37.
- Burrows, Paul, and Graham Loomes.** 1994. "The Impact of Fairness on Bargaining Behaviour." *Experimental Economics*, 21–41.
- Bursztyn, Leonardo, and Robert Jensen.** 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics*, 9: 131–153.
- Cadsby, C Bram, Ninghua Du, Fei Song, and Lan Yao.** 2015. "Promise Keeping, Relational Closeness, and Identifiability: An Experimental Investigation in China." *Journal of Behavioral and Experimental Economics*, 57: 120–133.
- Cain, Daylian M., Jason Dana, and George E. Newman.** 2014. "Giving versus Giving In." *Academy of Management Annals*, 8(1): 505–533.
- Camerer, Colin F.** 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.
- Cappelen, Alexander W., Trond Halvorsen, Erik Ø. Sørensen, and Bertil Tungodden.** 2017. "Face-saving or Fair-minded: What Motivates Moral Behavior?" *Journal of the European Economic Association*, 15(3): 540–557.
- Carlsson, Fredrik, Haoran He, and Peter Martinsson.** 2013. "Easy Come, Easy Go." *Experimental Economics*, 16(2): 190–207.
- Casal, Sandro, Werner Güth, Mofei Jia, and Matteo Ploner.** 2012. "Would You Mind if I Get More? An Experimental Study of the Envy Game." *Journal of Economic Behavior and Organization*, 84(3): 857–865.
- Charness, Gary, and Martin Dufwenberg.** 2006. "Promises and Partnership." *Econometrica*, 74(6): 1579–1601.
- Charness, Gary, and Martin Dufwenberg.** 2010. "Bare Promises: An Experiment." *Economics Letters*, 107(2): 281–283.
- Chaudhuri, Ananish.** 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics*, 14(1): 47–83.

- Chavez, Alex K., and Cristina Bicchieri.** 2013. "Third-Party Sanctioning and Compensation Behavior: Findings from the Ultimatum Game." *Journal of Economic Psychology*, 39: 268–277.
- Chen, Daniel L., Martin Schonger, and Chris Wickens.** 2016. "oTree – An Open-source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.
- Cherry, Todd L., and Jason F. Shogren.** 2008. "Self-Interest, Sympathy and the Origin of Endowments." *Economics Letters*, 101(1): 69–72.
- Cherry, Todd L., Peter Frykblom, and Jason F. Shogren.** 2002. "Hardnose the Dictator." *American Economic Review*, 92(4): 1218–1221.
- Cooper, David J., and John H. Kagel.** 2016. "Other-Regarding Preferences." *The Handbook of Experimental Economics*, 2: 217.
- Dana, Jason, Daylian M. Cain, and Robyn M. Dawes.** 2006. "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes*, 100(2): 193–201.
- Dana, Jason, Roberto A. Weber, and Jason Xi Kuang.** 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory*, 33(1): 67–80.
- Danilov, Anastasia, Christine Harbring, and Bernd Irlenbusch.** 2019. "Helping Under a Combination of Team and Tournament Incentives." *Journal of Economic Behavior & Organization*, 162: 120–135.
- Deck, Cary, Maroš Servátka, and Steven Tucker.** 2013. "An Examination of the Effect of Messages on Cooperation under Double-blind and Single-blind Payoff Procedures." *Experimental Economics*, 16(4): 597–607.
- DellaVigna, Stefano, John A. List, and Ulrike Malmendier.** 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1–56.
- Di Bartolomeo, Giovanni, Martin Dufwenberg, Stefano Papa, and Francesco Passarelli.** 2019. "Promises, Expectations & Causation." *Games and Economic Behavior*, 113: 137–146.

- Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman.** 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about others' Altruism." *American Economic Review*, 105(11): 3416–3442.
- Ederer, Florian, and Alexander Stremitzer.** 2017. "Promises and Expectations." *Games and Economic Behavior*, 106: 161–178.
- Ekström, Mathias.** 2012. "Do Watching Eyes Affect Charitable Giving? Evidence from a Field Experiment." *Experimental Economics*, 15(3): 530–546.
- Engel, Christoph.** 2011. "Dictator Games: A Meta Study." *Experimental Economics*, 14(4): 583–610.
- Exley, Christine L.** 2015. "Excusing Selfishness in Charitable Giving: The Role of Risk." *The Review of Economic Studies*, 83(2): 587–628.
- Fehr, Ernst, and Klaus M. Schmidt.** 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3): 817–868.
- Fehr, Ernst, and Simon Gächter.** 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–181.
- Fehr, Ernst, and Simon Gächter.** 2002. "Altruistic Punishment in Humans." *Nature*, 415(6868): 137.
- Fehr, Ernst, and Urs Fischbacher.** 2004. "Third-Party Punishment and Social Norms." *Evolution and human behavior*, 25(2): 63–87.
- Festinger, Leon.** 1962. *A Theory of Cognitive Dissonance*. Vol. 2, Stanford university press.
- Fischbacher, Urs.** 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10(2): 171–178.
- Fiske, Susan T.** 2018. *Social Beings: Core Motives in Social Psychology*. 4th edition, Hoboken, NJ : John Wiley & Sons, Inc.
- Forsythe, Robert, Joel L. Horowitz, Nathan E. Savin, and Martin Sefton.** 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*, 6(3): 347–369.

- Gächter, Simon, and Arno Riedl.** 2005. "Moral Property Rights in Bargaining with Infeasible Claims." *Management Science*, 51(2): 249–263.
- García-Gallego, Aurora, Nikolaos Georgantzis, and María J Ruiz-Martos.** 2019. "The Heaven Dictator Game: Costless Taking or Giving." *Journal of Behavioral and Experimental Economics*, 82: 101449.
- Gino, Francesca, Michael I. Norton, and Roberto A. Weber.** 2016. "Motivated Bayesians: Feeling Moral while Acting Egoistically." *Journal of Economic Perspectives*, 30(3): 189–212.
- Glasnapp, Douglas R., and John P. Poggio.** 1985. *Essentials of Statistical Analysis for the Behavioral Sciences*. CE Merrill Pub. Co.
- Gneezy, Uri, Agne Kajackaite, and Joel Sobel.** 2018. "Lying Aversion and the Size of the Lie." *American Economic Review*, 108(2): 419–53.
- Greenberg, Adam Eric, Paul Smeets, and Lilia Zhurakhovska.** 2015. "Promoting Truthful Communication through ex-post Disclosure." Available at SSRN: <https://ssrn.com/abstract=2544349>.
- Grossman, Zachary, and Joël J. Van der Weele.** 2017. "Self-Image and Willful Ignorance in Social Decisions." *Journal of the European Economic Association*, 15(1): 173–217.
- Güth, Werner.** 1994. "Distributive Justice – A Behavioral Theory and Empirical Evidence." In *Essays on Economic Psychology*, ed. Werner Güth and Hermann Brandstätter, 153–176. Springer-Verlag.
- Güth, Werner.** 2010. "The Generosity Game and Calibration of Inequity Aversion." *The Journal of Socio-Economics*, 39(2): 155 – 157.
- Güth, Werner, Kerstin Pull, Manfred Stadler, and Agnes Striebeck.** 2010. "Equity versus Efficiency? Evidence from Three-person Generosity Experiments." *Games*, 1(2): 89–102.
- Güth, Werner, M. Vittoria Levati, and Matteo Ploner.** 2012. "An Experimental Study of the Generosity Game." *Theory and Decision*, 72: 51–63.

- Haisley, Emily C., and Roberto A. Weber.** 2010. "Self-Serving Interpretations of Ambiguity in Other-Regarding Behavior." *Games and Economic Behavior*, 68(2): 614–625.
- Hamman, John R., George Loewenstein, and Roberto A. Weber.** 2010. "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship." *American Economic Review*, 100(4): 1826–1846.
- Hao, Li, and Daniel Houser.** 2017. "Perceptions, Intentions, and Cheating." *Journal of Economic Behavior & Organization*, 133: 52–73.
- Hobbes, Thomas.** 1651. *Leviathan*. Menston, Scolar P.
- Hoffman, Elizabeth, and Matthew L. Spitzer.** 1985. "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice." *The Journal of Legal Studies*, 14(2): 259–297.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith.** 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3): 653–660.
- Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith.** 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3): 346–380.
- Horton, John J., David G. Rand, and Richard J. Zeckhauser.** 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics*, 14(3): 399–425.
- Ismayilov, Huseyn, and Jan Potters.** 2016. "Why Do Promises Affect Trustworthiness, or Do They?" *Experimental Economics*, 19(2): 382–393.
- Jordan, Jillian, Katherine McAuliffe, and David Rand.** 2016. "The Effects of Endowment Size and Strategy Method on Third Party Punishment." *Experimental Economics*, 19(4): 741–763.
- Konow, James.** 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review*, 90(4): 1072–1091.
- Konow, James.** 2003. "Which is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature*, 41(4): 1188–1239.

- Kriss, Peter H., Roberto A. Weber, and Erte Xiao.** 2016. "Turning a Blind Eye, but Not the other Cheek: On the Robustness of Costly Punishment." *Journal of Economic Behavior & Organization*, 128: 159–177.
- Kurzban, Robert, Peter DeScioli, and Erin O'Brien.** 2007. "Audience Effects on Moralistic Punishment." *Evolution and Human Behavior*, 28(2): 75–84.
- Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber.** 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics*, 4(1): 136–163.
- LePine, Jeffrey A., Amir Erez, and Diane E. Johnson.** 2002. "The Nature and Dimensionality of Organizational Citizenship Behavior: a Critical Review and Meta-Analysis." *Journal of Applied Psychology*, 87(1): 52.
- List, John A.** 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy*, 115(3): 482–493.
- Lotz, Sebastian, Tyler G. Okimoto, Thomas Schlösser, and Detlef Fetchenhauer.** 2011. "Punitive versus Compensatory Reactions to Injustice: Emotional Antecedents to Third-Party Interventions." *Journal of Experimental Social Psychology*, 47(2): 477–480.
- Malmendier, Ulrike, Vera L. te Velde, and Roberto A. Weber.** 2014. "Rethinking Reciprocity." *Annual Review of Economics*, 6(1): 849–874.
- Mazar, Nina, and Chen-Bo Zhong.** 2010. "Do Green Products Make Us Better People?" *Psychological Science*, 21(4): 494–498.
- Mazar, Nina, On Amir, and Dan Ariely.** 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research*, 45(6): 633–644.
- Merritt, Anna C., Daniel A. Effron, and Benoît Monin.** 2010. "Moral Self-Licensing: When Being Good Frees Us to Be Bad." *Social and Personality Psychology Compass*, 4(5): 344–357.
- Mischkowski, Dorothee, Rebecca Stone, and Alexander Stremitzer.** 2019. "Promises, Expectations, and Social Cooperation." *The Journal of Law and Economics*, 62(4): 687–712.

- Mittone, Luigi, and Matteo Ploner.** 2012. "Asset Legitimacy and Distributive Justice in the Dictator Game: An Experimental Analysis." *Journal of Behavioral Decision Making*, 25(2): 135–142.
- Molenmaker, Welmer E., Erik W. de Kwaadsteniet, and Eric van Dijk.** 2016. "The Impact of Personal Responsibility on the (Un)willingness to Punish Non-Cooperation and Reward Cooperation." *Organizational Behavior and Human Decision Processes*, 134: 1–15.
- Nikiforakis, Nikos, and Helen Mitchell.** 2014. "Mixing the Carrots with the Sticks: Third Party Punishment and Reward." *Experimental Economics*, 17(1): 1–23.
- Oxoby, Robert J., and John Spraggon.** 2008. "Mine and Yours: Property Rights in Dictator Games." *Journal of Economic Behavior & Organization*, 65(3-4): 703–713.
- Pedersen, Eric J., Robert Kurzban, and Michael E. McCullough.** 2013. "Do Humans Really Punish Altruistically? A Closer Look." *Proceedings of the Royal Society B: Biological Sciences*, 280(1758): 20122723.
- Rege, Mari.** 2004. "Social Norms and Private Provision of Public Goods." *Journal of Public Economic Theory*, 6(1): 65–77.
- Rege, Mari, and Kjetil Telle.** 2004. "The Impact of Social Approval and Framing on Cooperation in Public Good Situations." *Journal of Public Economics*, 88(7-8): 1625–1644.
- Regner, Tobias.** 2018. "Reciprocity under Moral Wiggle Room: Is it a Preference or a Constraint?" *Experimental Economics*, 1–14.
- Rodriguez-Lara, Ismael, and Luis Moreno-Garrido.** 2012. "Self-Interest and Fairness: Self-Serving Choices of Justice Principles." *Experimental Economics*, 15(1): 158–175.
- Ruffle, Bradley J.** 1998. "More is Better, but Fair is Fair: Tipping in Dictator and Ultimatum Games." *Games and Economic Behavior*, 23(2): 247–265.

- Rutström, E. Elisabet, and Melonie B. Williams.** 2000. “Entitlements and Fairness: an Experimental Study of Distributive Preferences.” *Journal of Economic Behavior & Organization*, 43(1): 75–89.
- Schütte, Miriam, and Carmen Thoma.** 2014. “Promises and Image Concerns.” Munich Discussion Paper No, 2014–18.
- Schwartz, Steven, Eric Spires, and Rick Young.** 2019. “Why Do People Keep their Promises? A Further Investigation.” *Experimental Economics*, 22(2): 530–551.
- Selten, Reinhard.** 1978. “The Equity Principle in Economic Behavior.” In *Decision Theory and Social Ethics*. Vol. 17 of *Theory and Decision Library*, , ed. Hans W. Gottinger and Werner Leinfellner, 289–301. Springer Netherlands.
- Siegel, S., and N.J. Castellan.** 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.
- Soetevent, Adriaan R.** 2005. “Anonymity in Giving in a Natural Context — A Field Experiment in 30 Churches.” *Journal of Public Economics*, 89(11-12): 2301–2323.
- Tadelis, Steven.** 2011. “The Power of Shame and the Rationality of Trust.” Haas School of Business Working Paper 2011/3/2.
- Thaler, Richard H., and Cass R. Sunstein.** 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.
- Thunström, Linda, Todd L. Cherry, David M. McEvoy, and Jason F. Shogren.** 2016. “Endogenous Context in a Dictator Game.” *Journal of Behavioral and Experimental Economics*, 65: 117 – 120.
- Vanberg, Christoph.** 2008. “Why Do People Keep Their Promises? An Experimental Test of Two Explanations.” *Econometrica*, 76(6): 1467–1480.
- Van der Weele, Joël J., Julija Kulisa, Michael Kosfeld, and Guido Friebel.** 2014. “Resisting Moral Wiggle Room: How Robust is Reciprocal Behavior?” *American Economic Journal: Microeconomics*, 6(3): 256–264.