

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Essays on Non-Selfish Behavior

Permalink

<https://escholarship.org/uc/item/7gz5805w>

Author

Avalos Trujillo, Luis

Publication Date

2023

Peer reviewed|Thesis/dissertation

Essays on Non-Selfish Behavior

By

LUIS AVALOS-TRUJILLO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Economics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Anujit Chakraborty

Andrés Carvajal

Burkhard Schipper

Committee in Charge

2023

List of Figures

1.1 Behavioral comparison with regard to revealed expectations	14
2.3.1 Schematic representation of the real effort task.	41
2.3.2 Upstream reciprocity chain with payment details	42
2.5.1 Distribution of transfers according to treatment.	51
2.C.1 Demographic variables of subjects in the role of Worker 1	67
2.D.1 Overlay of clustering and principal component analysis	68
2.F.1 Sample of effort task graphic interface	74
2.F.2 Notice after receiving help.	75
2.F.3 Notice after receiving harm.	75
2.F.4 Comprehension test	78
2.F.5 Comprehension test for the treatment group: harm	80
3.3.1 Schematic representation of the real effort task.	89
3.3.2 Graphic interface used for the deliberation phase	90
3.3.3 Graphic interface used for the distribution phase	91

List of Tables

1.1	Summary of experimental interventions.	9
1.2	Donation by behavioral type.	12
1.3	Expectation preferences by type	13
1.4	Donation by behavioral type	16
1.5	Payment preference by type	16
1.6	Higher donation in the conditional task by type	17
2.5.1	Estimation results summary.	53
2.5.2	Interaction terms estimation summary.	56
2.5.3	Rates for different variables of interest in the reciprocity chain	57
2.5.4	Average marginal effects to explain the propagation rate	57
2.5.5	Comparison table.	58
2.A.1	Task sequence and monetary incentives for each role	64
2.C.1	Gender distribution by treatment group	66
2.D.1	Eigenvalue decomposition from Principal Component Analysis	68
2.D.2	Scoring coefficients from Principal Component Analysis	69
2.E.1	Results for the gratitude mechanism (equation 2.5)	70
2.E.2	Results for diff-in-diff estimation (equation 2.4)	71
3.3.1	Payoffs and surplus summary by task	89

Contents

Acknowledgements	vii
Preface	ix
1 Donations and Expectations	1
1.1 Introduction	1
1.2 The State of the Literature on Guilt Aversion	3
1.3 Theoretical Considerations and Hypotheses	6
1.4 Experimental Design	8
1.4.1 Experiment 1	9
1.4.2 Experiment 2	10
1.5 Results	11
1.5.1 Experiment 1	12
1.5.2 Experiment 2	15
1.6 Conclusion	17
Appendices	19
1.A Guilt Aversion Conformation Function	19
1.B Reference Dependent Conformation Function	21
1.C Functional Form Choices for Reference Dependent Conformation Function	24
1.C.1 The Linear Case	24
1.C.2 Square Root Case	25
1.C.3 Power Function	26
1.D Experimental Instructions	28
1.D.1 Experiment 1: Divider	28
1.D.2 Experiment 2: Divider	31
1.D.3 Volunteer	32

2	Upstream Reciprocity in the Battle of Good vs Evil	33
2.1	Introduction	33
2.2	Upstream Reciprocity in Biology, Psychology and Economics	37
2.3	Experimental Design	40
2.3.1	The Helping Task	42
2.3.2	The Harming Treatment	43
2.3.3	Alternating Tasks	43
2.3.4	Information Structure	44
2.4	Estimation Procedures	45
2.4.1	Survey Variables	45
2.4.2	Estimation Equations	47
2.5	Results	50
2.5.1	Control Group (Help)	51
2.5.2	Treatment Group (Harm)	52
2.5.3	Interaction Terms	54
2.5.4	Comparison across treatments	55
2.6	Conclusion and Discussion	59
	Appendices	62
2.A	Incentives and Tasks	62
2.B	Sample Size Determination	64
2.C	Demographics from Sample	66
2.D	Principal Component Analysis	66
2.E	Additional Estimation Results	68
2.E.1	Gratitude Mechanism	69
2.E.2	Difference-in-Difference	69
2.F	Transcripts of Experimental Procedure	70
2.F.1	Consent Form	70
2.F.2	PANAS	70
2.F.3	Social Capital Questionnaire	72
2.F.4	Economic Values Questionnaire	72
2.F.5	Effort Task	73
2.F.6	Demographic Questionnaire	76
2.F.7	Dictator Game	77
2.F.8	Dictator Game: Treatment Group	79
3	Morality as Determinant of Social Structure	81
3.1	Introduction	81

3.2 Discussion on Related Literature	85
3.3 Experimental Design	88
3.3.1 Production Phase	88
3.3.2 Deliberation Phase	90
3.3.3 Distribution Phase	91
3.3.4 Multiround Surplus Game	91
3.3.5 Additional Controls and Questionnaires	93
3.3.6 Summary of Experimental Procedure	94
3.4 Analysis and Estimation	95
3.4.1 Outcomes, Rules and Hypotheses	95
3.4.2 Proposed Estimation Procedures	97
3.5 Concluding Remarks	98
Appendices	100
3.A Moral Foundations Questionnaire	100
3.B Universalism Questionnaire	102

Acknowledgements

This dissertation crystalizes the effort of many years, and perhaps many generations. The completion of this work is as an act of service to all whose efforts made it possible.

My deepest gratitude to God who looked to this effort with mercy and allowed it to be completed, and my greatest devotion to the Blessed Virgin Mary whose example and assistance sustained me through the years that led to the completion of these essays. My deepest apology to all of those whom I have wronged in this journey.

This gratitude is extended to my parents and my family, for their great patience, love and support.

This essays owe a great deal to the mentorship and guidance of Anujit Chakraborty, Andrés Carvajal and Burkhard Schipper. I thank you for bearing with me and my sometimes unfeasible and eccentric ideas; I also want to thank you for zealously fulfilling your duty as advisors and as scientists. My formation as an Economist will forever bear your mark.

I also want to thank Arman Rezaee for guiding my exploration into development economics and for keeping my research approachable from outside the confines of experimental and behavioral economics. A special thanks to Gregory Clark because his daring investigations inspired me to pursue my own voice, and along with Andrés Carvajal, provided the financial support that allowed the completion of this project. This long journey was not without additional complications, my deepest gratitude to Katheryn Russ whose mentorship was, at times, therapeutic.

Last but not least, I would like to thank my colleagues, fellow students and friends who rode along this journey. For your friendship, advise and help, I thank you. Keisuke, Miki, Armando, Ninghui, Christina, Kalyani, Reem, Justin; I skip the formalities, this is also for you.

Preface

The essays in this dissertation are successive approximations to the understanding of the prosocial phenomenon in economic behavior. The longstanding paradigm is that of individual selfish maximization of utility, and behavioral economics has opened a new route to various forms of pro-sociality. In the present essays we explore social behavior when it is motivated by moral concerns.

In the first chapter, we investigate donation behavior and its dependence on expectations in the form of second order beliefs, which is colloquially known as guilt. The relationship is studied in a laboratory context through a modified third-party dictator game where the receiving end is a charitable organization and the third party is another experimental subject with a volunteering history, whose role in the game is to provide an expectation in the organization's behalf. The benchmark model in relation to second-order beliefs is guilt aversion, which is compared to other competing explanations found in the literature in addition to a model of reference dependent altruism. We are able to classify behavior according to each of the competing hypotheses. The results show strong evidence in favor of guilt aversion and some puzzling observations around the preference for expectations.

The second chapter is devoted to the study of upstream reciprocity, which is known colloquially as "pay-it-forward"; it is the act of reciprocating an act of kindness to an unknowing third party. In this chapter we propose an experimental measure of upstream reciprocity to enrich the concept of social or civic capital. It also contrasts upstream reciprocity to its evil counterpart: reciprocating an unkind act to an unsuspecting third party, which is termed "negative upstream reciprocity". The study combines an experimental intervention and a small survey from a general sample of the U.S. population. Results show that upstream reciprocity supplements trust: a previous positive social interaction makes a subject as prosocial as if they exhibit trust in strangers. The experiment allows for a contrast between positive and negative upstream reciprocity in the intensive margin, and a limited contrast in the

extensive margin. Results are amenable to a generalized principle of beneficence as outlined in Adam Smith's theory of moral sentiments.

The last chapter proposes an experimental intervention to study the effects of morality in the determination of social structure. It has been theorized that values, understood as process-regarding preferences, have important implications in institution formation. However, the question has remained largely unexplored. In this chapter we propose an experimental design to operationalize values through Haidt's moral foundations theory in an economic context. The experimental design focuses on a modified surplus division game which includes a steward figure who must perform the distribution of the surplus. The experiment attempts to measure the process through which outcomes are achieved rather than the outcomes themselves. The goal is to explore the differentiated social arrangements emerging from the pre-existing moral variability in a sample of college students in the United States.

Chapter 1

Donations and Expectations

LUIS AVALOS-TRUJILLO & ANUJIT CHAKRABORTY

1.1 Introduction

In recent years there has been an increased interest in the role of emotions in economic interactions. One of such approaches has focused on the role of guilt. In economics, the venture was pioneered by the introduction of the concept of guilt-aversion (Charness and Dufwenberg, 2006), which rests on the theoretical framework of psychological games (Geanakoplos et al., 1989). In psychology, the concept of guilt is understood as the unpleasant emotions that an individual suffers associated with possible objections to his or her actions, inaction, circumstances or intentions (Baumeister et al., 1994). It has also been referred to as the feeling of distress that arises whenever someone has done harm to somebody else, when someone receives more than what they deserve or when a moral standard has been violated, even when “nobody is harmed or disappointed or knows about the incident” (Baumeister et al., 1994). On the other hand, guilt-aversion is a more specific concept that occurs within the context of a monetary transfer. The receiving party holds an expectation about how much will be transferred, such expectation is a second order belief to the party making the transfer. Guilt in this context is derived from the perceived “harm” that is inflicted on the counterpart by failing to act according to the expectation.

However, experimental results on guilt aversion are not conclusive and experience shows how difficult it is to identify empirically. Guilt aversion depends on

Declared Exempt by UC Davis IRB Administration ID: 1591145-1. This research project was made possible thanks to the Financial support of Andrés Carvajal.

beliefs from others' anticipations which are difficult to manipulate. On one hand, some studies have found evidence in favor of guilt aversion (Dufwenberg and Gneezy (2000); Charness and Dufwenberg (2006); Bacharach et al. (2007); Dufwenberg et al. (2011); Bellemare et al. (2018)). Some of those studies relied on finding a positive correlation between transfers and second order beliefs. The approach was not exempt of possible confounds, and changes to the experimental design led to the opposite conclusion: rejecting the correlation between transfers and second order beliefs (Vanberg, 2008; Ellingsen et al., 2010; Kawagoe and Narita, 2014). Some other designs led to alternative explanations on similar experimental settings like preferences for surprise-seeking (Khalmetzki et al., 2015) or hump-shaped behavior (Balafoutas and Fornwagner, 2017).

The present research intends to contribute to that literature by means of the following proposals. First, the introduction of a context which could prove to be meaningful to guilt aversion. While most of the literature on guilt-aversion had focused on experimental instances of trust games and dictator games, the present research utilizes the dictator game in the context of a donation within an experimental environment. The donation context intends to provide a situation where a moral standard is salient. In the proposed experimental setup, a player will divide an endowment between herself and a charitable institution chosen from a menu.

Second, we introduce a third-party dictator game where the recipient of the transfer is different from the expectation-holding party. In previous experiments the recipient and the expectation-holding party were the same player. If the expectation-holding party and the recipient were the same player, positive relationship between the expectations and the amount transferred, as predicated by guilt-aversion, would be indistinguishable from the behavior observed by an agent whose preferences are altruistic but reference dependent on the expectations. By utilizing the expectations provided by a third-party, we make sure that the positive relation cannot be interpreted as reference dependence.

Hence, our experimental setup consists of a player who, after choosing a charitable institution from a menu, has to perform two donation tasks: first, a direct donation (a regular dictator game); and second, a strategy-method dictator game, whose transfers are conditional on the expectations provided by the third-party. The treatments are predicated on the order of the tasks and on the revelation (or lack thereof) of the expectations from the third-party. Once accounting for the heterogeneity from the motivations to donate, the experiment is able to identify a substantial portion

of guilt-averse subjects. It is important to note that the third-party is never made aware of the contributions made to the charity, and that the donor is aware of this fact. The design contributes to identifying guilt even though “nobody is harmed or disappointed or knows about the incident”, as predicated in the psychology literature.

Third, the present research deepens our understanding of how guilt-averse subjects behave in regards to expectations. If expectations have an option to be revealed, guilt-type subjects prefer to reveal the expectations prior to making the unconditional donation. For those subjects, the information contained in the expectations influences their donation with the objective of decreasing their guilt. On the other hand, when expectations do not have an option to be revealed, guilt-type subjects prefer to avoid making choices in reference to expectations. The observations are puzzling and some of the explanations proposed are a preference for “moral wiggle-room”, motivated avoidance or procedural preferences.

The rest of the paper is structured as follows: section 1.2 presents a review on the experimental literature related on guilt-aversion. In section 1.2 the theoretical framework is briefly presented as well as the research hypotheses. In section 1.4 the experimental design is outlined and results are shown in section 1.5. Conclusions are left to section 1.6

1.2 The State of the Literature on Guilt Aversion

In Economics, the study of guilt aversion has been constrained to the study of a two person interaction in which one party is either transferring a monetary amount to a passive recipient (a dictator game) or a two person interaction in a trust game where both parties are deciding upon monetary outcomes. The phenomena is called guilt aversion and can be loosely defined as the experience of disutility that stems from failing to act according to the perceived expectation from the other player, where expectation in this context is an action that would lead to a monetary transfer to the counterpart. Failure to act according to such expectation would produce a lower psychological payoff. Guilt in this context is derived from the perceived “harm” that is inflicted on the counterpart by failing to act according to the expectation. Guilt aversion was advanced by Charness and Dufwenberg (2006) and later formalized by Battigalli and Dufwenberg (2007).

Experimental results on guilt aversion have found mixed results and experience shows how difficult it is to identify empirically. Guilt aversion depends on beliefs from

others' anticipations which are difficult to manipulate. Usually, they are elicited by asking subjects about what they believe other players are expecting (Guerra and Zizzo, 2004; Bacharach et al., 2007). In the original paper where guilt aversion was introduced, Charness and Dufwenberg (2006) present an experiment designed to test the effect of guilt aversion in a trust game with hidden action, which abstracts a principal-agent relationship (the trustor/principal receives an amount M which sends to the trustee/agent, he in turn returns an amount which is multiplied by a positive integer). The authors show that second order beliefs of the trustee, measured from self-reports, were correlated with a higher returned amount.

However, their experiments could suffer from several confounds. First, it could be that the trustees' actions are anchored on their own guess from the expectations of the trustor. Another confound might be that trustees think that trustors are also thinking like them, by a consensus effect (Ross et al., 1977), causing that, in the aggregate, trustees sending large back-transfers are exactly those that believe the trustor is expecting a large back-transfer.

In order to provide a more stringent test for guilt aversion, free from consensus effects, Ellingsen et al. (2010) introduced a novel methodological contribution to the study of second order beliefs. Instead of eliciting them directly, they ask the trustor for their belief about the back-transfer, this information is then passed (unknowingly) to the trustee. This covert transfer of information is key to avoid strategic revelation from the trustor and it fixes the second order beliefs of the trustee. Besides the hidden action trust game, a similar approach is used in a dictator game and a regular trust game. In all their experiments no correlation between beliefs and transfers is found. Beyond the debate of the appropriateness of their experimental setup, their results represent the major challenge to the guilt aversion model.

More recently Khalmetski et al. (2015) presented an experiment based upon the design in Ellingsen et al. (2010) to prove their intuition that the lack of correlation was due to the desire to surprise their counterpart by exceeding their expectations, specially in the dictator game. Their argument is that if there are enough number of people willing to exceed the expectations from the recipient in a dictator game, then the aggregate observation should show no correlation because the surprise seeking agent's actions are negatively correlated with the expectations, while guilt averse agents' actions are positively correlated with expectations. To proof their argument they design a dictator game but they elicit behavior by means of the strategy method (Selten, 1967; Mitzkewitz and Nagel, 1993; Brandts and Charness, 2011). Their re-

sults show evidence for surprise-seeking behavior which dilutes the evidence for guilt averse behavior in the aggregate, among other results which confirm the presence of consensus effects. This study was the first to point out that the heterogeneous motivations behind the monetary transfer were behind the negative results found in previous studies using correlations.

A replication study by [Balafoutas and Fornwagner \(2017\)](#) further supports that explanation by classifying the multiple motivations behind the transfer by using the strategy method as in [\(Khalmetzki et al., 2015\)](#) in a dictator game. Subjects were classified as selfish if they transferred zero for all levels of expectations, as unconditional altruists if they transferred a positive amount irrespective of expectations, as guilt averse if they increased their donation according to expectations, surprise seeking if the transfers were negatively correlated with expectations. Additionally, they provide an additional classification for the subjects who describe an optimal transfer function that is “hump-shaped”, that means their donations increase with expectations up to a threshold and they negatively correlate with expectations above the threshold. Classification was made possible by the use of the Pearson correlation coefficient and in a strategy method with high granularity. In their results, they did not find a prevalent behavioral type but rather the subjects are more or less evenly distributed amongst the types except for the unconditional altruist, under which a minor proportion of subjects is classified. The prevalence of hump shaped optimal transfer functions is a salient feature from their results. Although the authors do not conclude that the cause of unkind behavior for higher expectations is due to a desire to punish the recipient, they clearly point out a limit to guilt aversion behavior.

It is worth mentioning that in the original study by [Charness and Dufwenberg \(2006\)](#) there is a treatment with pre-play communication. Although our experiment does not consider pre-play communication, it is worth mentioning that some of the studies allowing for it have also showed evidence against guilt-aversion. Some of the earlier experiments showed a preference for promise keeping per-se, undermining the preference for fulfilling expectations (not letting down others) [\(Vanberg 2008\)](#), while in a more recent study [\(Kawagoe and Narita, 2014\)](#), the concept of guilt aversion is extended to aversion to unfulfilled expectations that were provoked by the first-party, mimicking more closely an aversion to breaking promises. The authors find no correlation between expectations and transfers in neither the original nor the extended concept.

1.3 Theoretical Considerations and Hypotheses

The first objective of the experimental design consists in uncovering the presence or absence of guilt-averse subjects. Given the substantial evidence against guilt-aversion, we did not hold a position *a priori* regarding its presence. Recalling the model of guilt aversion is given by

$$\max_t u(M - t) - \theta \max\{0, e - t\}$$

where M is the endowed amount, t is the transfer to be effected by the subject and e is the expectation in the form of second order beliefs. In the existing literature, e represents the second order belief from the subject about how much the recipient is expecting and θ is a sensitivity parameter. From the current model, it is evident that a testable implication of guilt-aversion consists on an optimal transfer function that is non-decreasing in e , hence the use of the strategy method to identify guilt-averse subjects. Appendix [1.A](#) shows a proof under certain technical conditions.

Given the conflicting evidence from [Balafoutas and Fornwagner \(2017\)](#) and [Khalmetzki et al. \(2015\)](#), where they both study the optimal transfer functions via a strategy method, our initial hypothesis is that subjects with non-decreasing optimal transfer functions will be a minority of subjects within the sample.

It is worth mentioning that the presence of a non-decreasing optimal transfer function cannot be readily interpreted as guilt-aversion. This is particularly problematic if we assume that the subject has altruistic preferences with preference dependence with regards to expectations in the style of [Kőszegi and Rabin \(2006\)](#). Such case was initially analyzed by [Breitmoser and Tan \(2013\)](#) and can be adapted to our setting by means of the following expression governing the subject's preferences

$$\max_t u(M - t) + \alpha \mu(e, t)$$

where the reference dependent function $\mu(e, t)$ takes the form

$$\mu(e, t) = \begin{cases} u_2^G(t - e) & t \geq e \\ \lambda u_2^L(e - t) & t < e \end{cases}$$

where $u_2^G(t - e)$ and $u_2^L(e - t)$ represent the utility of the counterpart in the gains and loss domain respectively and λ represents the loss aversion parameter. Depending on

the choice of functional forms, such model could explain the hump-shaped optimal transfer function. However, it is possible to arrive at a testable implication under certain technical conditions. Namely under continuity and twice differentiability of the utility function, in addition to the regular assumptions for the reference dependent component, it can be shown that the optimal transfer will never be above the expectation level e . The proof and discussion is reserved to appendix [1.B](#). A couple of examples of different specifications for the reference dependence conformation function can be found in appendix [1.C](#).

In order to mitigate such concerns, the present experiment considers decoupling the recipient of the transfers from the expectation holding party. Then, if the experimental subject's optimal transfer function shows some relationship to the expectations of a party different from the recipient, it is difficult to interpret such behavior as altruistic towards the expectation holding party when the latter is ignorant of the outcome, and the subject is aware of it. Note that the subject could still be altruistic towards the recipient; but in such case, the transfer should be independent from the expectations of a third-party. Expectations provided in such way could be interpreted as a proxy for the second order beliefs. Alternatively, they could be interpreted as an extension to the current guilt aversion model by pointing to a concept of guilt that is derived from the standards of conduct provided by a third party. For simplicity, we will refer to the expectation-holding party as the third-party.

Finally, we are also interested in the preferences that subjects have towards expectations. Few studies have analyzed a similar situation, among the most relevant is [Dana et al. \(2006\)](#) where found that subjects would prefer to pay to avoid a dictator game or [Dana et al. \(2007\)](#) where subjects decide to acquire information regarding the payoffs to the other participant or could keep those payoffs hidden allowing for the subject to act self-interestedly ("moral wiggle-room"). Our goal is to extend the notion of guilt aversion by uncovering a preference for the expectations themselves. On one hand, knowing the expectations can allow the guilt-averse subjects to accommodate their behavior to expectations. On the other hand, it could be the case that having those expectations being made explicit is unpleasant in itself, because they negatively arouse the guilt averse subject to comply. Our initial stance was agnostic with regards to the subject's behavior with regards to their preference for expectations.

1.4 Experimental Design

The experiment consists of two active players: the dictator and the third-party; and one passive player, a charitable institution. The dictator has to split a monetary amount between herself and the charity by means of two tasks: 1) a conditional task, consisting of a strategy method dictator game based on the guess from the third-party; and 2) an unconditional task, which consists of a regular dictator game. Each dictator is matched to a third party, whose task is to estimate the donation from the matched dictator in the unconditional task. In the context of the experiment, the expectations from the third-party were referred to as a “guess”. Data from the third-parties is collected first. The choices from the dictator are private and they were made aware of this fact.

Incentivizing the guess from the third-party and ensuring the privacy of the choices from the dictator presented an additional complication. Note that if we had paid for an accurate guess the volunteer could have inferred the donation. The solution we implemented was to recruit additional subjects in the dictator role but under a public condition. In the public condition, we stated that their donations could be made known to the third-party. After recruiting all the players in the third-party role, they were randomly matched to the dictators, and only the those matched to dictators in the public condition were paid an additional amount if their guess was correct. Players in the third-party role matched in the private condition were not eligible for an additional payment. Such complicated arrangement was presented to them as: “A randomly selected group of participants can win an additional \$ X if their guess is accurate” where the amount X varied depending on the experiment. It is to be noted that all of the analysis is carried out only with the subjects under the private condition.

The experimental treatments varied the order of the tasks as will be described later. Previous to their tasks, subjects in the dictator role had to choose a charitable institution from a menu consisting on American Cancer Society, Citizen Schools, American Red Cross, Doctors without borders and Clean Water Foundation. The menu was the same in all experimental treatments.

Since a majority of subjects in this experiment were recruited from Amazon’s Mechanical Turk (MTurk), it is important to note that all experiments contained one or more quizzes. Only subjects who passed each quiz were allowed to continue. Failing any quiz resulted in an immediate disqualification from the study. All the sample

	I	II
Endowment	\$5	\$10
Granularity	1-5 point-wise	6 Intervals
Volunteer Composition	MTurkers only	College & MTurkers
Donor Composition	Mturkers	Mturkers
Order of Tasks	1. Conditional 2. Expectation Preference 3. Unconditional	1. Unconditional 2. Conditional 3. Payment Preference

Table 1.1: Summary of experimental interventions.

sizes reported refer to those subjects who passed the quizzes and hence completed the study.

1.4.1 Experiment 1

A total of 115 subject pairs were recruited from Amazon’s Mechanical Turk (MTurk), 19 of them in the public condition and 96 in the private condition. Dictators received a participation fee of \$1 and \$5 to split between herself a a charity. Subjects in the third-party role received \$2 and the chance to win an additional \$2 for providing a correct guess.

Subjects in the dictator role proceed in the following way: first, they select a charity from the menu and then they were briefed about their tasks. Afterwards, they performed a quiz to test their understanding of the tasks and to filter out automated responses. If they failed to answer 4 out of 6 questions correctly they were disqualified from the study and ineligible for payment. After passing the quiz, subjects performed the conditional task (a strategy method dictator game where the donation had to be specified for a guess of: \$0, \$1, \$2, \$3, \$4 or \$5). Before proceeding to the unconditional task (regular dictator game) the subject was asked to choose when they would like to see the guess from the volunteer: before or after performing the unconditional task. Considering their choices, either the guess was revealed first and then the regular dictator game was performed, or the regular dictator game was performed followed by the revelation of the guess. Finally, subjects answered a demographic questionnaire. Payment was carried out from either task to be selected at random.

The subject pool for subjects in the third-party role requires some further expla-

nation. A majority of subjects in the third-party role were MTurkers who responded positively to the question: “Have you volunteered to charity in the past two years”. For that reason, subjects in the third-party role were referred to as “volunteers” during the experiment. The sample of volunteers was selected from an initial sampling of 148 MTurk participants. From there 67 declared to have recently volunteered to a charity while 30 more were randomly selected from the pool who declared negatively. The reason for using MTurks who have recently volunteered was to provide meaning to their guess and mimic a solicitation scenario. Subjects in the dictator role were only informed that “a majority of subjects in the volunteer role had recently volunteered for a charity” and that the volunteers had already been briefed about the dictator’s task and that they had already provided their guess.

1.4.2 Experiment 2

A total of 80 subject pairs were recruited for this experiment, 10 in the public condition and 70 in the private condition. All of the subjects in the dictator role were recruited from Amazon’s Mechanical Turk. 26 subjects in the third-party role were recruited from MTurk while 44 of them were recruited from undergraduate charity clubs at the University of California, Davis. Third-party subjects from MTurk were not pre-screened and the volunteer portion of the sample was comprised of the college students from the charity clubs. Again, subjects in the dictator role were informed of the volunteer, but their choices remained private in the private condition. Dictators were informed that “a majority of subjects in the volunteer role are currently volunteering at community-service driven college clubs” ¹

Dictators received a participation fee of \$1 and \$10 to split between herself and the charity selected from the menu. Subjects in the third-party role received \$2.50 as participation fee and could earn an additional \$2.50 through the mechanism described earlier.

For subjects in the dictator role, the experiment proceeded in the following manner. First, they selected a charity from a menu. Afterwards, subjects were briefed

¹Initially, all of the subjects in the third-party role were going to be recruited from charity clubs at UC Davis. The disruption of college activities due to COVID-19 forced us to continue the experiment by recruiting the third-party subjects from MTurk. Given that dictators are only briefly informed of the composition of the subject pool, we considered that this feature is not instrumental to the results obtained in this experiment.

about the unconditional task followed by a quiz. Only subjects that answered correctly 2 out of 4 questions were allowed to continue. Afterwards, they performed the unconditional task. Later, they were briefed about the conditional task, followed by a quiz. Only subjects that answered 2 out of 3 questions correctly were allowed to continue. After the second quiz, subjects performed the conditional task. The strategy-method in the conditional task is based on the following intervals for the guess from the volunteer: \$0, \$1-\$2, \$3-\$4, \$5, \$6-\$7 and \$8 or more. Finally, subjects were given the choice to influence their payment. They were asked to choose under which task they would prefer to be paid: the conditional task or the unconditional task. Whichever task they selected had an 80% chance to be used for payment.

1.5 Results

In interpreting the results, we follow the classification strategy by [Balafoutas and Fornwagner \(2017\)](#) and classify all subjects according to their behavior in the conditional task.

1. Selfish type, for individuals who donated zero for every expectation.
2. Unconditional Altruists, for individuals who donated a constant positive amount for every expectation.
3. Guilt averse, for individuals who are not classified as selfish or altruists and whose optimal transfer function is non-decreasing in expectations.
4. Surprise seeking, for individuals who were not classified as selfish or altruist but whose optimal transfer function is non-increasing in expectations.
5. Hump Shape, for individuals who exhibited an increasing trend in their donations up to a certain expectation. For expectations above or equal to that level, donations exhibited a decreasing trend. Also, if donations were increasing and decreased at the highest expectation range.
6. Other, for any individual who could not be classified in the above types. Most of the individuals classified in this group exhibited noisy observations around a certain level of donations.

Type	Effective Donation	Number	Percentage
Selfish***	\$0	19	20%
Unconditional Altruists	\$1.17	6	6%
Guilt Averse	\$1.92	27	28%
Surprise Seeking	\$2.23	4	4%
Hump Shape	\$1.30	8	8%
Other**	\$2.90	32	33%
		96	100%

Table 1.2: Donation by behavioral type. Pairwise difference in means vs guilt-type are computed (p-value *** < 0.001, ** < 0.05, * < 0.1)

1.5.1 Experiment 1

Behavioral classification and the average donation in each type can be observed in table 1.2. Note that the majority of subjects are classified as “other”, while the second largest classification is guilt averse. The high prevalence of guilt-averse subjects must be noted. Most subjects that cannot be classified exhibit noisy positive donations, which usually results as an artifact from the strategy method. This effect is considerably more pronounced since the subject pool is recruited from MTurk, the interface used was based on sliders, and that the experiment was for small stakes. In the follow-up experiment we addressed some of these issues.

The first surprising result was the small proportion of subjects classified as hump-shaped. Given that the expectation-holding party is never made aware of the outcome, it is not surprising that there are very few subjects classified as surprise-seeking.

Average donation in the unconditional task amounted to 1.92 (0.168)², meanwhile in the conditional task the effective donation averages 1.66 (0.157). Effective donation is defined as the donation carried out once the guess from the matched volunteer is considered. An alternate metric could be the estimated expected donation when the distribution of guesses is considered. In such case, the expected donation is 1.61 (0.141). Note that donations are higher in the unconditional task.

The preference for expectations is summarized in table 1.3. We observe no global preference about revealing expectations. However, 80% (22 out of 27) of the guilt

²All the numbers in parenthesis are standard errors

Type	Order Choice			
	Task First		Expectation First	
Selfish	8	20%	11	20%
Unconditional Altruists	3	7%	3	5%
Guilt Averse	5	12%	22	40%
Surprise Seeking	2	5%	2	4%
Hump Shape	5	12%	3	5%
Other	18	44%	14	25%
	41	100%	55	100%

Table 1.3: Expectation preferences by type. Fisher exact test: difference in behavior is significant at 4.4%

averse subjects indicate a preference to reveal expectations prior to making a donation. Also note that all other types lack of a clear preference for the order of the tasks.

It is interesting to investigate how the guilt-types behave with respect to the expectation that was revealed. Behavior is summarized in figure [1.1](#). The figure considers only the observations from the subjects who chose to reveal the guess. The figure is based on the following definitions: extended guilt and information offset. Extended guilt refers to the difference between the expectations and the transfer. Positive extended guilt means that the subject “disappointed” her counterpart by transferring an amount below the expectation, hence the positive guilt. Negative extended guilt corresponds to the case of the transfer is above expectations, which can be interpreted as a surprise. Information offset is the difference between the amount donated in the unconditional task and the amount donated in the conditional task at the level of expectation that was revealed to the subject. Positive offset means that the donation in the unconditional task, after seeing the guess, is higher than their donation on the conditional task with respect to the same guess.

Out of the 96 subjects, 55 preferred to reveal the guess prior to the unconditional donation. Of those 55, 22 had been previously classified as guilt-type and 33 were not. In figure [1.1a](#) we can see that guilt-types show a tendency to react to the revealed expectation. Note that zero guilt is attained when the subject matches the expectation that was just revealed. Subjects that were classified as non-guilt, seem to show a disregard to the revealed expectation as can be seen in figure [1.1b](#), where the distribution of extended guilt is widely spread across the range. Non-guilt types exhibit a more consistent behavior across tasks as summarized in the informa-

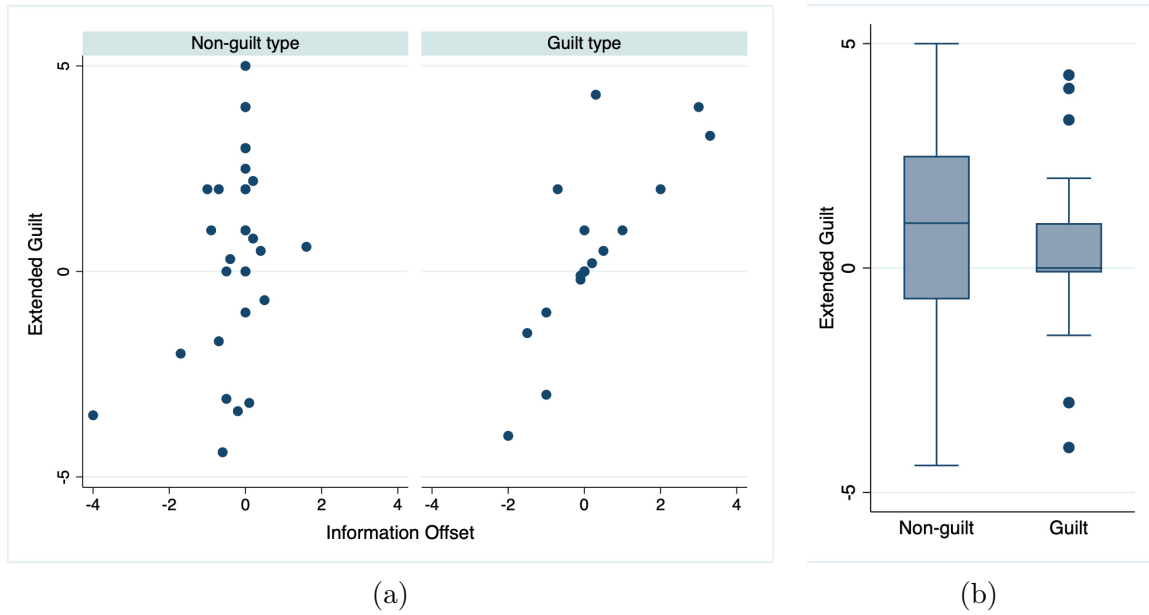


Figure 1.1: Behavioral comparison with regard to revealed expectations. Extended guilt is defined as expectation minus transfer. Information offset is the donation after seeing the expectation minus the donation chosen during the conditional task. F test: positive slope on “Guilt type” is significant (p-value<0.001).

tion offset axis, as most subjects are clustered around the zero. On the other hand, subjects classified as guilt, transfer an amount that is close to the expectation just revealed as can be seen more clearly in figure 1.1b, where most of the observations are clustered around zero guilt. Also, they tend to revise their previous donation in the conditional task by getting it closer to the observed expectation as can be inferred from the positive slope in figure 1.1b

Interestingly, when we pool the 55 subjects who saw the expectation prior to the unconditional donation, and compute the Pearson correlation coefficient between the revealed expectation and the transfer we conclude that the correlation is not significantly different from zero (coefficient of -0.06 with a p -value of 0.66). This result confirms the observations in Khalmetski et al. (2015) where the heterogeneity in the observed behavior in relation to expectations, obscured the results in the aggregate. This implies that the zero correlation results obtained in Ellingsen et al. (2010) could be due to heterogenous responses to the second order beliefs.

1.5.2 Experiment 2

Given the prevalence of noisy observations in the previous experiment, we adjusted the stakes to the upside offering \$10 in the dictator game and allowing the strategy method to be played in the ranges \$0, \$1-\$2, \$3-\$4, \$5, \$6-\$7 and \$8 or more. Also, recall that the order of the tasks is reverted: the unconditional task is performed first and later the conditional task. At the end of the two tasks and prior to the demographic questionnaire, subjects manifested under which task to receive payment, the selected task will be chosen for payment with 80% chance.

Average donations amounted to 3.79 (0.380) in the unconditional task and 2.66 (0.330) in the conditional task when the expectation from the matched volunteer is considered and 2.68 (0.294) when the distribution of expectations is used to compute the expected donation. Note again that average donations are higher in the unconditional task.

Behavioral classification of subjects are shown in table 1.4. Observe that most subjects are classified as guilt-types, such prevalence had not been reported in the literature so far. Higher stakes, while keeping constant the granularity of the strategy method, helped reduce the amount of noisy observations. However they still represent a considerable proportion of subjects within the sample.

Type	Effective Donation	Number	Percentage
Selfish**	\$0	14	20%
Unconditional Altruists*	\$4.44	9	13%
Guilt Averse	\$2.75	27	39%
Surprise Seeking	\$6.50	1	1%
Hump Shape	\$3.00	4	6%
Other***	\$5.61	15	21%
		70	100%

Table 1.4: Donation by behavioral type. Pairwise difference in means vs guilt-type are computed (p-value *** < 0.001 , ** < 0.05 , * < 0.1).

Type	Payment Preference			
	Unconditional	Conditional	Unconditional	Conditional
Selfish	12	27%	2	8%
Unconditional Altruists	7	16%	2	8%
Guilt Averse	18	41%	9	35%
Surprise seeking	0	0%	1	4%
Hump Shape	1	2%	3	12%
Other	6	14%	9	35%
	44	100%	26	100%

Table 1.5: Payment preference by type. Fisher exact test: difference in behavior is significant at 3.1%.

Preferences for payment are shown in table [1.5](#). In this case we observe an aggregate preference to be paid under the unconditional task, guilt-types are also part of this trend. The choice is paradoxical for the guilt-types since in the previous experiment they had expressed a preference to reveal the expectations prior to the unconditional task. Furthermore, since the strategy method provides a perfect “hedge” for any of the volunteer’s guesses, and since the strategy method came second in the list, it was to be expected that guilt-averse subjects would prefer to be paid under the conditional task.

Such behavior is difficult to interpret, in one hand it could be explained as a choice for “moral wiggle-room”. In this case, subjects would prefer to make an unconditional donation to have an excuse to fail to meet expectations. However, donations

Higher Donation in the Unconditional Task			
Type	Count	Percentage	Avg. Difference
Selfish	1	7%	0.07
Unconditional Altruists	3	33%	0.22
Guilt Averse	21	78%	1.93
Surprise Seeking	0	0%	-2.50
Hump Shape	4	100%	2.86
Other	10	67%	1.03

Table 1.6: Number of subjects with higher donation in the conditional task by type.

in the unconditional task are higher than in the conditional task for most of the guilt-types (see table 1.6), such explanation would require subjects to hold beliefs on the higher end of the spectrum. Another possible explanation could be motivated avoidance. Under this explanation, subjects would prefer to avoid expectations to avoid the distress from having to consider a donation in reference to them, and facing the choice of not meeting them. Given that the last experiment showed that guilt-type subjects can be swayed by the expectation, motivated avoidance remains a possibility. Lastly, such behavior could be explained by a preference for procedure. This would entail that subjects prefer to make a direct donation rather than having to donate in reference to the expectations from the third-party, this could be interpreted as a distaste for the strategy method.

1.6 Conclusion

The present experiments showed the prevalence of guilt aversion behavior when we consider guilt aversion as the disutility emanating from unfulfilled expectations in the form of second order beliefs. In this case, we showed that those second order beliefs are relevant in the decision making process even when their source is a third-party who is ignorant of the outcome. The experiments point that expectations matter even when no one is harmed or nobody will know about the incident. However, since the expectation-holding party is not the recipient of the transfer, the guilt-aversion presented here constitutes an extension to the present understanding of guilt-aversion.

We also conclude that subjects with guilt-type responses react differently to the

expectations from the third-party. If expectations are always revealed, guilt-type subjects prefer to reveal the expectations prior to making the unconditional donation. For those subjects, the information contained in the expectations influences their donation. On the other hand, when expectations are never revealed, guilt-type subjects prefer to avoid making choices in reference to expectations. The observations are paradoxical and some of the explanations proposed are a preference for “moral wiggle-room”, motivated avoidance or procedural preferences.

Our results also highlight the importance of the proper identification techniques in the presence of heterogeneous behavior. We showed that via correlation coefficients it is not an appropriate method to identify guilt aversion. Meanwhile, by means of a classifier, guilt-aversion was identified and its differential behavior was assessed.

Appendix

1.A Guilt Aversion Conformation Function

The model of guilt aversion was originally introduced by [Charness and Dufwenberg \(2006\)](#) and later formalized by [Battigalli and Dufwenberg \(2007\)](#). At the core of the model lies the second order belief of the player about his own actions. This means that the player infers what his opponent is expecting from him and then he derives a disutility if he fails to deliver as expected. In the context of the model, actions correspond to a monetary transfer. Hence, the player will feel a disutility proportional to the difference between the second-order belief and the actual transfer, this difference is called guilt. In their original paper, guilt aversion was experienced only when the party receiving the transfer was also the one forming expectations. In our environment, we extend this notion to situation where the party forming expectations is different from the party receiving the material transfer.

Formally, a guilt aversion conformation function will take the form

$$v = -\max\{0, z(e - t)\}$$

This guilt function is a slight generalization from [\(Charness and Dufwenberg, 2006\)](#), in which an additional function z evaluates the distance to the second order belief. The function z can be thought of as a truncated gain-loss function, akin to the reference dependent utility. In this case, the reference is given by the expectation e . Given the definition of the disappointment function, z corresponds only to the loss side of a reference dependence utility, of which we assume the usual properties on z' and z'' namely $z' > 0$ and $z'' \leq 0$.

The guilt function enters the sender's optimization problem as

$$\begin{aligned} \max_t U(D = 1, M - t, t - e) &= u(M - t) - \theta \max\{0, z(e - t)\} \\ \text{subject to } 0 &\leq t \leq M \end{aligned} \tag{1.1}$$

Assume u is twice differentiable and $u' > 0$ and $u'' \leq 0$.

Proposition 1. *Under a guilt aversion conformation function, optimal transfers are non-decreasing in e .*

Proof. The proof considers the following two cases.

I. Linear u and z

In this case assume $u(M-t) = M-t$ and $z = -\max\{0, e-t\}$ which corresponds to the standard assumptions in [Charness and Dufwenberg \(2006\)](#). The problem to be solved is

$$\max_t M - t - \theta(e - t)$$

whose solution is $t^* = e$ if $\theta \geq 1$ and $t^* = 0$ if $\theta < 1$. The optimal transfer is increasing in e whenever $\theta > 1$ and zero whenever $\theta < 1$.

II. General u and linear z

In this case assume $u(M-t)$ with $u' > 0$, $u'' < 0$ and $z = -\max\{0, e-t\}$. The problem to be solved is

$$\max_t u(M-t) - \theta \max\{0, e-t\}$$

whose solution is $t^* = e$ or t^* solves $\theta = u'(M-t)$. The optimal transfer is either increasing in e or independent of it.

III. General Case

The easy case is when $v = 0$, which corresponds to the optimal solution $t^* = e$. Note that it is never optimal to transfer above e since it will diminish the selfish utility term $u(M-t)$ while keeping the v term constant at zero. The interesting case is when $v > 0$, which would arise when the optimal transfer is strictly below e , in other words we are exploring the domain $t \in (0, e)$ assuming a solution exists. If $t^* = 0$, a local disturbance in e will not produce an effect unless the restriction $t \geq 0$ is binding exactly at zero.

In $(0, e)$ first order conditions are given by,

$$-u'(M-t) + \theta z'(e-t) = 0. \tag{1.2}$$

Again, the solution t^* exists and is continuous by Berge's theorem. Again, the implicit function theorem allows to conclude $t^*(e)$ is differentiable and we can take the derivative of equation [1.2](#) with respect to e to get

$$[\theta v''(e-t^*) - u''(M-t^*)] \frac{dt^*}{de} = \theta z''(e-t^*). \tag{1.3}$$

Since $z'' > 0$ because we are in the loss domain, and by concavity of u the term in brackets in the left hand side is positive (i.e. $[\theta z''(e - t^*) - u''(M - t^*)] > 0$). Finally, we notice that the right hand side is also positive, hence $\frac{dt^*}{de} > 0$. \square

The spirit of this proposition is that if the optimal transfer is $t^* = e$, then its derivative with respect to e is positive and the same would be true if $0 < t^* < e$. The derivative will also be positive when the restriction $t \geq 0$ is binding exactly at zero. In other cases the derivative will be zero. Observe that it is never optimal to transfer anything above e and that the optimal transfer is increasing in the second-order belief.

1.B Reference Dependent Conformation Function

One alternative explanation to the observed behavior would be that of reference dependent structure with regards to expectations. Under such model, the dictator would give targeting a reference point, which in this case would be given by her second order beliefs.

The first assumption to be made is that the reference dependent structure follows the assumptions in [Kőszegi and Rabin \(2006\)](#). This feature tries to capture the notion that the sender might be thinking that the other player is expecting a certain amount, either because such amount would be considered “just” or “fair” ([Rabin, 1993](#); [Charness and Rabin, 2002](#)), or because it is the norm within the given context ([Kimbrough and Vostroknutov, 2019](#)). The problem of the donor is given by

$$\begin{aligned} \max_t U(D = 1, M - t, t - e) &= u(M - t) + \theta\mu(t - e) \\ \text{subject to } 0 &\leq t \leq M \end{aligned} \tag{1.4}$$

To begin formalizing our discussion, first assume the following properties.

Assumption 1. *The selfish component u is continuous, twice differentiable, strictly increasing and concave.*

On the other hand, the reference dependent component μ is a “gain-loss” function which evaluates a payoff y according to

$$\mu(y; r) \equiv \mu(y - r) \tag{1.5}$$

In the reference dependence literature μ is usually defined as $\mu(m(y) - m(r))$ where m is a usual consumption utility assumed to be differentiable and strictly increasing. In the present paper we are assuming m is the identity function in order to simplify the analysis (i.e. $m(x) = x$ for all x). This simplification does not cause much harm since we are mostly interested in evaluating gains and losses against a reference. Using the function m adds an additional step without aiding to the explanation of the phenomena that concerns us.

Assumption 2. *The properties of function μ are assumed to be as follows:*

- i. $\mu(x)$ is continuous for all x , differentiable for $x \neq 0$ and $v(0) = 0$*
- ii. $\mu(x)$ is strictly increasing*
- iii. If $0 < x < y$, then $\mu(y) + \mu(-y) < \mu(x) + \mu(-x)$*
- iv. $\mu''(x) \leq 0$ for $x > 0$ and $\mu''(x) \geq 0$ for $x < 0$*
- v. $\mu'_-(0)/\mu'_+(0) \equiv \lambda > 1$ where μ'_- is the left hand derivative and μ'_+ the right hand derivative.*

These assumptions correspond to the formalization by [Bowman et al. \(1999\)](#), which abstract the stated and unstated assumptions of prospect theory's value function ([Tversky and Kahneman, 1979](#)). Loss aversion in small stakes is captured by (ii) while loss aversion in large stakes is captured by (iv). Diminishing sensitivity is captured by assumption (iii). The interpretation of the “gain-loss” function in the current setting is that of a subject who experiences a disutility for non-conformation and a utility for going beyond conformation (transferring more than what was expected). Conforming (i.e. transferring as expected) does not provide additional utility. The subject is risk averse when the transfer is below expectations, which means that the subject is worried about delivering even lower. When the transfer is above the expectation, the subject becomes risk loving which means that he is willing to incur in further deviations above the expected amount. Such behavior would be characteristic of an individual who is guilt averse when delivering below expectations but surprise seeking when delivering above expectations. The loss aversion assumptions ensure that disappointing looms larger than surprising.

To simplify the analysis, we will make use of the following redefinition of the “gain-loss” function

$$v(x) = \begin{cases} G(x), & x \geq 0 \\ -L(-x), & x < 0 \end{cases}$$

Notice that in strict sense we are defining G and L from v , and hence some of its properties are inherited like continuity and twice differentiability. Also note that both G and L are increasing and concave.

Observations in pilot studies and the data from [Khalmetski et al. \(2015\)](#) show a prevalence for transfers that are not equal to the reference point and also different from zero. Therefore it is of vital importance to be able to characterize optimal transfers that are interior (i.e. excluding optimal transfers equal to zero, e or M) as much as possible and also to be able to determine their behavior as a function of expectations.

In order to find such characterization, we require an additional assumption that relates risk aversion between the reference dependent component and the self-regarding component. It is reasonable to assume that when considering risky prospects the subject might feel more concerned with respect to his own material payments than the psychological penalty derived from non-conformation. In other words, we assume risk aversion is lower with respect to non-conformation (i.e. the losses side of the reference dependent component) than the material utility from her payoff. If we consider the disutility of non-conformation to be an interpersonal phenomena as in the experimental psychology literature, the disutility of non-conforming is evoked due to the perceived failure to act according to a standard or the perceived harm inflicted into the expectation holding party. The latter interpretation is well supported in economic experimental literature that has found that loss aversion with respect to losses to someone else is lower than the loss aversion with respect to the subject's own payoff ([Polman, 2012](#); [Andersson et al., 2016](#); [Füllbrunn and Luhan, 2017](#)).

Then we assume that, on the loss domain, absolute risk aversion is greater for the payments directed to the psychological penalty from non-conformation.

Assumption 3. *If $t < e$, then*

$$\frac{u''(M-t)}{u'(M-t)} < \frac{L''(e-t)}{L'(e-t)}$$

Under such assumption we arrive to the following proposition.

Proposition 1. *Under assumptions 1, 2 and [3](#), the optimal solution t^* satisfies $t^* \notin (0, e)$.*

Proof. We will proceed by showing that in the loss domain $(0, e)$ there are no local

maxima.

First order conditions reveal the possibility of an interior critical point whenever there exists a $t \in (0, e)$ that satisfies

$$\theta = \frac{u'(M - t)}{L'(e - t)}.$$

We will denote this point as t_L^* .

However, this critical value corresponds to a local minimum. In order to prove that, take the second order derivative on problem [1.4](#) to obtain

$$U''(t) = u''(M - t) - \theta L''(e - t)$$

From first order conditions we get

$$\theta = \frac{u'(M - t_L^*)}{L'(r - t_L^*)},$$

by using assumption [3](#) and using the fact that $L''(r - t) < 0$ and $u'(M - t) > 0$,

$$\frac{u''(M - t_L^*)}{L''(r - t_L^*)} > \frac{u'(M - t_L^*)}{L'(r - t_L^*)}.$$

Therefore

$$\frac{u''(M - t_L^*)}{L''(r - t_L^*)} > \theta$$

hence $U''(t_L^*) > 0$. □

1.C Functional Form Choices for Reference Dependent Conformation Function

1.C.1 The Linear Case

The linear case corresponds to all functional forms being linear yielding the following expression for the utility function

$$U(D = 1, M - t, t - e) = M - t + \theta\mu(x - e)$$

where $\mu(x - e) = x - e$ when $x > r$ and $\mu(x - e) = -\lambda(e - t)$ when $t \leq e$.

It is very easy to show that if $\alpha > 1$, the optimal transfer is $t^* = M$. Now for the case of $\alpha \leq 1$ there are two cases: if $\theta\lambda > 1$ then the optimal transfer is $t^* = e$; else if $\alpha\lambda \leq 1$ then $t^* = 0$.

1.C.2 Square Root Case

The special case $\sigma = \frac{1}{2}$ greatly simplifies the computations and it is very useful to illustrate the mechanisms of the model. In this case, the selfish component is given by $u(M - t) = \sqrt{M - t}$; meanwhile, the other regarding component is

$$\mu(t - r) = \begin{cases} \frac{1}{2}\sqrt{t - e} & , t \geq e \\ -\frac{\lambda}{2}\sqrt{e - t} & , t < e \end{cases}$$

Again, analysis on the gains domain is straightforward since $U(t)$ will be conformed by the addition of two concave functions, therefore any critical point will be a local maximum. From first order conditions we obtain an expression for the local maximum

$$t_G^* = \frac{r + \theta^2 M}{1 + \theta^2}.$$

On the losses side, first order conditions imply

$$\theta\lambda = \sqrt{\frac{e - t}{M - t}},$$

and we denote as t_L^* as the value of t that solves the previous equation. From second order conditions we find that $U''(t) > 0$ if and only if

$$\theta\lambda > \left(\frac{e - t}{M - t}\right)^{3/2}.$$

Observe

$$\left(\frac{e - t}{M - t}\right)^{3/2} < \sqrt{\frac{e - t}{M - t}} < \sqrt{\frac{e}{M}}$$

Therefore if $\theta\lambda > \sqrt{\frac{e}{M}}$ then $U'(t) > 0$ for all $t \in [0, r]$, which implies t_G^* is the global optimum. If $\theta\lambda < \sqrt{\frac{e}{M}}$ then $U(t)$ is increasing in some portion of $[0, e]$ and decreasing in another. From first order conditions we can see that $\theta\lambda > \left(\frac{e - t_L^*}{M - t_L^*}\right)^{3/2}$,

hence the critical value in $[0, e]$ correspond to a global minimum. Again, this implies that the only possible solutions are $t_L^* = 0$ or $t_L^* = e$. The former can be readily discarded since $U'_+(e) > 0$, while the latter could be a global maximum if

$$A^{**} \equiv \sqrt{1 + \theta^2} \sqrt{1 - \frac{e}{M}} + \alpha \lambda \sqrt{\frac{e}{M}} < 1.$$

The previous discussion is summarized as follows. Whenever u and μ are utility functions of the power family and $\sigma = 1/2$, the global optimum t^* is described as

$$t^* = \begin{cases} t_G^* & \text{if } \alpha \lambda > \sqrt{\frac{e}{M}} \\ t_G^* & \text{if } \alpha \lambda < \sqrt{\frac{e}{M}} \text{ and } A^{**} > 1 \\ 0 & \text{if } \alpha \lambda < \sqrt{\frac{e}{M}} \text{ and } A^{**} < 1 \end{cases}$$

1.C.3 Power Function

Consider then a model where both u and μ are from the power function family, hence

$$u(M - t) = \frac{(M - t)^{1-\sigma}}{1 - \sigma}$$

and

$$\mu(t - e) = \begin{cases} \frac{1}{1-\sigma}(t - e)^{1-\sigma} & t \geq e \\ -\frac{\lambda}{1-\sigma}(e - t)^{1-\sigma} & t < e \end{cases}$$

given $\sigma < 1$.

To analyze optimal behavior under such model we need to proceed by cases. In the gains domain (i.e. $t > e$), analysis is straight forward since both u and μ are concave in $[r, M]$, therefore any critical point will be local maximum. The critical point is obtained by first order conditions and it is given by

$$t_G^* = \frac{\theta^{1/\sigma} M + e}{1 + \theta^{1/\sigma}}.$$

However, on the losses side, analysis is not straightforward and certain combination of parameters will not make t_G^* the global optimum. This is intuitively true since a very low value of α will prompt the subject to neglect the payments to the other agent and transfer zero. To formalize this notion first observe that the first

order conditions for $t < e$ are

$$\left(\frac{e-t}{M-t}\right)^\sigma = \theta\lambda \quad (1.6)$$

which implies that $\mu'(t) > 0$ whenever $\left(\frac{e-t}{M-t}\right)^\sigma < \theta\lambda$.

Also note $\frac{e-t}{M-t}$ is decreasing in t hence

$$\left(\frac{e-t}{M-t}\right)^\sigma < \left(\frac{e}{M}\right)^\sigma$$

Now consider two cases, first if $\theta\lambda > \left(\frac{e}{M}\right)^\sigma$ then $\mu(t)$ is increasing in $[0, e]$. The selfish component $u(t)$ is also increasing in $[0, e]$, which by continuity of $U(t)$ and the fact that U is increasing in $[e, t_G^*]$ imply that $U(t_G^*) > U(t)$ for any $t \in [0, t_G^*]$, hence the global optimum is t_G^* .

The second case is if $\theta\lambda < \left(\frac{e}{M}\right)^\sigma$ then v is increasing only in a subset of $[0, e]$. It can be shown that the critical point t_L^* is a local minimum, since $\mu''(t) < 0$ if and only if $\theta\lambda < \left(\frac{e-t}{M-t}\right)^{1+\sigma}$ and since

$$\left(\frac{e-t_L^*}{M-t_L^*}\right)^{1+\sigma} < \left(\frac{e-t_L^*}{M-t_L^*}\right)^\sigma = \theta\lambda$$

we conclude $\mu''(t_L^*) > 0$, which means t_L^* is a local minimum. Note that the last equality follows from first order conditions (equation [1.6](#)).

The previous analysis shows that the only candidates for a local maximum in the loss domain are either $t_L^* = 0$ or $t_L^* = r$. The latter can be readily discarded as a global maximum since $U'_+(r) > 0$, to see this observe that $\mu'_+(r) = \infty$. On the other hand $t_L^* = 0$ will be global maximum if and only if $U(0) > U(t_G^*)$, which occurs whenever

$$A^* \equiv \theta\lambda \left(\frac{r}{M}\right)^{1-\sigma} + (1 + \theta^{1/\sigma}) \left(1 - \frac{e}{M}\right)^{1-\sigma} < 1$$

The previous discussion can be summarized in the following proposition.

Proposition 2. *Whenever u and v are utility functions of the power family, the global optimum t^* is described as follows*

$$t^* = \begin{cases} t_G^* & \text{if } \alpha\lambda > \left(\frac{e}{M}\right)^\sigma \\ t_G^* & \text{if } \alpha\lambda < \left(\frac{e}{M}\right)^\sigma \text{ and } A^* > 1 \\ 0 & \text{if } \alpha\lambda < \left(\frac{e}{M}\right)^\sigma \text{ and } A^* < 1 \end{cases}$$

Notice that in order to obtain a hump-shaped optimal transfer function, the parameters α and σ should be sufficiently low for any given λ , such that for high levels of the reference e we will get $A^* < 1$ yielding an optimal transfer of zero.

1.D Experimental Instructions

1.D.1 Experiment 1: Divider

[Consent for omitted for brevity]

First, select a charitable organization that appeals to you. [A menu appears showing the names, logos and brief descriptions of the following charities:]

- American Cancer Society.- Nationwide voluntary health organization dedicated to eliminating cancer
- Citizen Schools.- American nonprofit organization that partners with middle schools across the United States to expand the learning day for children in low-income communities.
- American Red Cross.- Humanitarian organization that provides emergency assistance, disaster relief, and disaster preparedness education in the United States.
- Doctors without borders.- International humanitarian medical non-governmental organization of French origin best known for its projects in conflict zones and in countries affected by endemic diseases.
- Clean Water Fund.- American environmental advocacy group the group that focuses on canvassing and gaining support for political issues and candidates, on issues related to water.

[New page]

You have been given \$5. Your job is to decide how much of the \$5 dollars you want to donate to [charity selected previously].

You will do so through two tasks: the Conditional Task and the Unconditional Task. You have to complete both tasks, but only one of the two tasks will be selected randomly for your payment.

In each task you can donate any amount between \$0 and \$5, and keep the rest for yourself. Your total income from this survey will be equal to \$1 plus the amount of money you keep for yourself in the randomly selected task.

[New page]

You have been randomly matched to another individual, who will be called the Volunteer throughout this survey. Volunteers have already completed a different survey previously. A majority of individuals in the volunteer role are other MTurkers who recently volunteered for a charity.

Your earnings will be transferred to you within three weeks of your participation date.

The total amount allocated to [charity selected previously], by you and other participants, will be calculated and then donated on your behalf.

You have 24 hours to complete this HIT.

[New page]

CONDITIONAL TASK - INSTRUCTIONS

We have given the Volunteers the following information about you. We refer to you as "the Divider" in third-person: "We, the surveyors, will give the Divider \$5, then she will be asked to split the \$5 between herself and a charitable organization she can choose from a menu. She will be paid the amount that she decides to keep to herself and we will donate the rest on her behalf." And we have asked the Volunteer "How much do you believe the Divider is going to donate?" allowing her to respond with a number between 0 and 5, and we have recorded her response in our system.

Now, you have to decide how much to donate to [charity selected previously], but

you do this conditional on the guess provided by the Volunteer. In other words, you to make your donation choice for a range of possible expectations from the Volunteer.

This will be immediately clear if you take a look at the following figure.

[A figure showing 5 sliders is shown. Each slider correspond to a hypothetical guess from the volunteer]

For example, if the volunteer guessed that you were going to donate \$2, your response in the slider for a guess of \$2 will count towards the final payment.

[Another figure showing 5 sliders is shown. One slider is highlighted indicated a possible guess with the associated donation.]

In the next page you will have an opportunity to practice for the Conditional Task.

[The practice round has the same interface as the real task but it is followed by a small quiz asking what would be the payoff to the subject under two hypothetical scenarios using the subject's responses in the practice round. Four additional comprehension questions are included in the quiz. Two questions refer to the task of the volunteer, one asks who will be the recipient of the donation, and the final question refers to when the subject should expect to receive their payment.]

Task 1. Conditional Task

Out of the \$5, indicate how much will you donate for a range of possible expectations from the Volunteer.

Please note that your donation choices are confidential, and no one, not even the matched Volunteer will be informed of your choice.

[Slider interface for conditional donations is shown.]

The final task is to choose a donation to [charity selected previously].

[Expectation Reveal Question:]

Before proceeding to the final task, choose when would you like to see the Volunteer's guess:

First perform task, then see the guess. (1)

First see the guess, then perform task. (2)

Task 2. Unconditional Task

This is your final task.

Choose your donation to [charity selected previously].

Please note that your donation choices are confidential, and no one, not even the matched Volunteer will be informed of your choice.

[A single slider is shown for unconditional donation]

[Note that the conditional task can appear after or before showing the guess from the volunteer as chosen by the subject.]

1.D.2 Experiment 2: Divider

[The main interface for experiment 2 is shared with experiment 1 except in the order of the tasks. In experiment 2, the unconditional task comes first, proceeded by a quiz. The conditional task comes last and it is preceded by a small quiz to test the comprehension of the conditional task. The expectation reveal question does not appear in this experiment. Also, the experiment runs with \$10 instead of \$5, and the sliders are kept at 6 under the following ranges: \$0, \$1-\$2, \$3-\$4, \$5, \$6-\$7 and \$8 or more.]

[Experiment 2 concludes with the payment choice:]

You have the option to choose your payment system from the two choices below.

The payment system on the left assigns a significantly higher chance (80%) of paying you based on your Unconditional choice. The payment system on the right assigns a significantly higher chance (80%) of paying you based on your Conditional choice (where you made your choice conditional on Volunteer's expectations).

The Unconditional choice helps you maintain the same donation (and payment) irrespective of the Volunteer's actual expectations. The conditional choice helps you

adjust your donation (and payment) to the Volunteer's actual expectations.

[The following options are presented]

Unconditional Choice Preferred

- **Unconditional Choice - 80% chance**
- Conditional Choice - 20% chance

Conditional Choice Preferred

- Unconditional Choice - 20% chance
- **Conditional Choice - 80% chance**

1.D.3 Volunteer

[Volunteer surveys are identical except for the monetary amounts. In experiment 1 it is \$5 and in experiment 2 it is \$10.]

You will be paired randomly with another participant from Amazon's Mechanical Turk. We will refer to her as the Divider, with female pronouns. We, the surveyors, will give the Divider \$5, then she will be asked to split the \$5 between herself and a charitable organization she can choose from a menu. She will be paid the amount that she decides to keep to herself and we will donate the rest on her behalf.

Your task consists in answering the question "How much of the \$5 do you expect the MTurker to donate?" and a short survey.

Your earnings will be transferred to you within three weeks of your participation date. All your answers to this survey are anonymous.

How much of the \$5 do you expect the Divider to donate? A randomly selected group of participants can win an additional \$2 if their guess is accurate. We will consider your guess to be accurate if it is within $\pm\$0.50$ of the actual donation choice

[A slider interface is provided to indicate a guess.]

Chapter 2

Upstream Reciprocity in the Battle of Good vs Evil

2.1 Introduction

Reciprocity is a very well-known phenomenon in Economics. It is well understood both theoretically (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) and empirically (Falk et al., 2008; Guala, 2010; Cabral et al., 2014). The concept is based upon the principle of “I help you if you help me”. However, a less studied phenomenon is indirect reciprocity, which is based on the principle “I help you and somebody else will help me”.

Indirect reciprocity comes in two types: downstream and upstream. Downstream reciprocity can be described as: B helps C, and then B receives help from A. The psychological dynamics of downstream reciprocity stem from the social image created by B after helping A. This particular social dynamic has also been referred to as “image scoring” and has been substantially studied in both psychology and biology (Nowak and Sigmund, 1998; Wedekind and Milinski, 2000; Ule et al., 2009).

Upstream reciprocity, on the other hand, means: A helps B, and in turn B helps C. In other words, the recipient of an altruistic interaction pays it forward to a novel individual, creating a new cooperative relationship between B and C. Contrary to the case of downstream reciprocity, there are few studies focusing on upstream reci-

Declared Exempt by UC Davis IRB Administration ID: 1867182-1. This research was carried out with the financial support from the Dearing Funds from the Economics Department at UC Davis. Special thanks to Professor Gregory Clark for his support to the completion of this project.

procuity.

In psychology, there is a scant but substantive literature which places gratitude as the main mechanism behind upstream reciprocity (Bartlett and DeSteno, 2006; DeSteno et al., 2010; Tsang, 2006, 2007). However, the concept has received very little attention in the economics literature except in few mentions in studies related to “image scoring” (Dufwenberg et al., 2001; Bolton et al., 2005) and applications to public good provision (Greiner and Levati, 2005; Steiger et al., 2014) and intergroup conflict (Hugh-Jones et al., 2019).

The objective of the present research is to enrich our notion of social or civic capital through the introduction of an experimental measure of upstream reciprocity. Social capital has been defined as “the persistent and shared values and beliefs that help a group overcome the free-rider problem in the pursuit of socially valuable activities” (Guiso et al., 2011). Unfortunately, there is no consensus as to which are those persistent shared values except for trust and trustworthiness, which have been shown to be measurable at the laboratory level and at the survey level. More importantly, trust has been shown to have important correlation with macroeconomic outcomes like economic development (Algan and Cahuc, 2010), financial development (Guiso et al., 2004) and international trade (Guiso et al., 2009).

Upstream reciprocity is an ideal concept to enrich our notion of social capital since theoretical research in social biology has shown that it is a promoter of the evolution of cooperative behavior, aiding the evolution of direct reciprocity (Nowak and Roch, 2007). Upstream reciprocity can be used as a measure of the propensity to propagate prosocial actions, measurable at the individual or group level. Behaviors that could fall under prosocial actions range from workplace interactions (Chancellor et al., 2018a, b) to adoption and propagation of norms in the societal scale, specially through the mechanism of horizontal transmission (Boyd and Richerson, 1988).

On the other hand, negative upstream reciprocity can be used as a measure of the propagation of anti-social actions like criminal behavior. Having a measure for the propagation of anti-social actions could inform estimates of *upstream* criminality given an initial mass of antisocial behavior. Another possible application lies in the study of intergroup conflict, since it is known that violence committed to one group is retaliated to any member from the perpetrator’s group (Chagnon, 1988; Haushofer et al., 2010; Horowitz, 1985; Hugh-Jones et al., 2019; Shayo and Zussman, 2011). A formal measurement could be used to study if a group with higher *in-group* negative

upstream reciprocity exhibits higher *out-group* negative reciprocity. Such measures could be carried out experimentally or through a survey question following the example set by the Global Preference Survey (Falk et al., 2018).

The contributions of the present study are manifold: the proposed experiment is the first incentivized measure of the propagation of prosocial action, free of income effects and free of deception, in a sample drawn from a general population in the United States. Furthermore, this study contrasts the propagation of prosocial actions with anti-social (or self-seeking) actions, which we will term *negative upstream reciprocity*. To the knowledge of the author, the proposed intervention is the first experimental measure of negative upstream reciprocity across the literatures in psychology, social biology and economics.

In addition to the experimental intervention, the social capital questionnaire from the World Values Survey (WVS) (Haerpfer et al., 2020) is used to uncover the relationship between upstream reciprocity and trust. The economic values questionnaire from WVS is used to uncover the relationship between upstream reciprocity and policy preferences around welfare, government ownership, competition and the dilemma of equality vs equity. The present study is the first to uncover the relationship between upstream reciprocity, social capital and economic values.

The experiment consists of a reciprocity chain, where four subjects interact in sequence. The first player can start the reciprocity chain by having a kind act towards the second. The second player is the main subject of study who chooses to perform a kind act to an unknowing third player. The second player must perform a tedious task (a series of on-screen tasks similar to CAPTCHAs). The first player has the option to partially relieve her from that burden by paying a monetary amount. After finishing her task, the second player can reciprocate to a third player through a dictator game, i.e., the second player decides how much to transfer to the third player from an additional endowment.

In the treatment group, the first player has the option to increase her own payoff at the expense of the second player, whose task becomes more burdensome. Afterwards, the second player is able to reciprocate this self-seeking action by engaging in a dictator game with a third player, and the second player is given the additional option to “steal” from the third player up to a pre-specified amount.

Upstream reciprocity is shown to supplement trust, enhancing prosocial behavior.

In other words, subjects who report a lack of trust in others pay-forward in similar magnitude as trusting subjects after they experience an act of kindness from the previous player. Subjects who identify with more progressive policies also pay-forward in higher magnitudes, in line with previous experimental literature in progressiveness and altruistic giving (Clark et al., 2017; Dawes et al., 2012; Fosgaard et al., 2019; Gilens and Thal, 2018; Kerschbamer and Müller, 2020). It is interesting to note that *negative upstream reciprocity* is shown to have no correlation with any explanatory variables except having received harm in a previous stage. This leaves us ignorant at which are the psychological mechanisms underlying the propagation of harmful actions.

In addition, results show that kind or unkind actions are passed forward at approximately equal rates (extensive margin), results are supported through simple difference in means tests and logit regression. A comparison in the intensive margin is limited by the experimental design considering that the magnitude of the gift paid forward might be contingent on the dictator endowment and the transfer range. Difference in means tests and difference-in-difference analysis are used to discern the magnitude of these effects. Results show that positive and negative upstream reciprocity have similar magnitudes, with the caveat that results are inconclusive due to the large size of the standard errors.

In relation to the psychology literature and unorthodox approaches in economics, we find that gratitude plays a role in reciprocating kindness, supporting the previous findings in psychology. The results are amenable to the parsimonious explanation provided by Adam Smith’s Theory of Moral Sentiments and the principles of beneficence and harm. In this case, the results show that upstream reciprocity is a generalization of the principles: “beneficence leads to gratitude, which leads to reward.” and “harm leads to resentment, which leads to punishment” (Smith and Wilson, 2017, 2019a). This study makes it clear that the reward or punishment can be paid back to any member of society, not exclusively to whomever realized the initial act of beneficence or harm.

The rest of the paper is structured as follows: in section 2.2 we present the literature on upstream reciprocity from various fields. The most prominent are to be found in psychology, while social biologists have posed important contributions. In economics, the mentions are scant and most of them do so tangentially. In section 2.3 we discuss the experimental design including tasks and questionnaires to be completed by the subjects. In section 2.4 we describe the estimation procedures and the

results are shown in section 2.5. Conclusion and further discussion is reserved to section 2.6

2.2 Upstream Reciprocity in Biology, Psychology and Economics

Early literature on upstream reciprocity is mostly theoretical since it is associated with the discussions around the mechanisms for the evolution of cooperation. The term “upstream reciprocity” was coined by Boyd and Richerson (1989) while proposing a mechanism that could sustain large scale cooperation in human groups. Through simulation studies, Nowak and Roch (2007) proposed that the evolution of indirect (upstream) reciprocity can promote cooperation if it is linked to a mechanism for the evolution of cooperation, in this case the mechanism is the possibility of direct reciprocity.

Upstream reciprocity was first identified experimentally by economists, who encountered it as a limiting case of “image scoring” when the information on reputation was not available (Dufwenberg et al. 2001 Bolton et al. 2005). When no reputation information is available and agents interact in a chain, engaging in altruistic behavior can be thought of initiating and continuing upstream reciprocity. Both studies found evidence in favor of upstream reciprocity, however their results are not conceptualized as such. Cyclical chains were also studied in economics in early research on indirect reciprocity (Greiner and Levati 2005).

Experimental studies in psychology have focused on uncovering the mechanisms behind upstream reciprocity. The seminal work in psychology is by Bartlett and DeSteno (2006) who identified the connections between gratitude and upstream reciprocity. Even though the mechanisms of upstream reciprocity are not the objective of this study, it is important to provide a brief overview since among the possible explanations, psychologists have almost reached a consensus on gratitude being the main driver. The primary alternative explanation is *affect*, which argues for changes in the helping attitude derive from a change in mood. Another alternative explanation is a norm of upstream reciprocity, which would entail that humans have internalized that reciprocating is socially desirable, even when reciprocating to a third party. Finally, and perhaps more akin to economic reasoning, some have postulated that upstream reciprocity operates through a belief channel.

Although mood alterations have been found to induce helping (Berkowitz and Daniels, 1964; Isen and Levin, 1972), singling out mood alterations as the sole cause for upstream reciprocity is highly questioned. One common mechanism to tease apart the observed behavior is to induce a mood alteration by chance or through someone’s generosity. Several studies strongly suggest the presence of gratitude due to the intentional help received and find no evidence in favor of the affect theory (Bartlett and DeSteno, 2006; DeSteno et al., 2010). Further studies that do not rely on self reports but rather on observed behavior on scripted dictator games, support the gratitude mechanism in sure and risky outcomes. (Tsang, 2006, 2007).

Another conjecture is that upstream reciprocity operates through a belief channel (Romano et al., 2021). Such claim has not been tested in psychology, however recent results in Economics have brought the question into the table (Schwerter and Zimmermann, 2020; Buckenmaier and Dimant, 2021). Given the legacy of belief literature in Economics it is fitting to conduct a rigorous study of this question, but testing such hypothesis lies beyond the scope of the present study.

Lastly, it has been conjectured that a norm of reciprocity could be the underlying mechanism of upstream reciprocity. Such explanation is problematic since it is commonly held that norms emerge from behavior, and not behavior from norms (Opp, 1982; Sen and Airiau, 2007). However, the evidence on a norm for upstream reciprocity should not be disregarded altogether since anecdotal evidence suggest a strong social pressure to continue pay-it-forward chains at drive throughs.¹ However it is open to discussion if a norm of upstream reciprocity exists in other contexts. Since the norm of upstream reciprocity presents theoretical and experimental issues that are substantially different from the present research, we consider that its study merits a separate investigation.

One of the most prominent experimental studies in upstream reciprocity consists in a comparison between upstream and downstream reciprocity by Stanca (2009). In their experiments, subjects are paired in groups of 4 where they play a trust game. The base case has 2 players engaged in a regular trust game. The treatments consist of changing the recipient of the back-transfer to a third player (upstream), or having the back-transfer from a third player come after the initial transfer (downstream).

¹It is interesting to note the numerous newspaper columns, internet forums and Q&A websites containing discussions about breaking pay-it-forward chains. Despite some occasional opposition there seems to be a social norm to continue pay-it-forward chains at drive throughs

They conclude that the initial transfer is higher under upstream reciprocity treatment than downstream reciprocity and even direct reciprocity. Also, the back-transfers are higher under upstream reciprocity when measured by the strategy method.

However, the forces of upstream reciprocity seem to be faint in comparison to other mechanisms, specially in the presence of downstream reciprocity. In a relatively recent field experiment, [van Apeldoorn and Schram \(2016\)](#) did not find any evidence in favor of upstream reciprocity. Their study involved an online community of travelers and hosts where a costly service can be provided for free and information about each member past actions is available. In other words, each traveller can request service from hosts, and the latter provide their services for free. However, roles can reverse: a host can become a traveler anytime and request service from other hosts. Note that we are using the terms “traveler” and “host” although it is not clear what the service is because it was kept undisclosed. All members can see the past actions from any member making a request for service. Requests can be accepted or declined based on past history. It is no surprise that the study found strong evidence in favor of downstream reciprocity but no evidence of upstream reciprocity. However, without the presence of direct reciprocation or reputation information, upstream reciprocity has been found to be prominent. In the field experiment conducted by [Mujcic and Leibbrandt \(2018\)](#), they showed that in a traffic environment, subjects are more than twice as likely to act generously and stop after someone else has stopped for them.

In recent literature, developmental psychologists have traced the emergence of upstream reciprocity to children between 3 and 4 years old ([Beeler-Duden and Vaish, 2020](#)). In economics, recent articles have brought the issue of upstream reciprocity to the forefront. A recent study by [Steiger et al. \(2014\)](#) exposes the connections between indirect reciprocity and public good provision; their study validates the relevance of upstream reciprocity in a context of prosocial action with economic implications.

More recently, [Schwerter and Zimmermann \(2020\)](#) designed an experiment in which subjects are the recipients of a dictator game prior to playing a trust game. The dictator can be played by a computer or by a human being. The authors found that willingness to trust is substantially higher after a positive social experience; i.e. the subject sends more in the trust game when receiving a positive amount from a human being than from the computer. They conjecture that beliefs change after a social experience and corroborate their views through an additional experimental treatment with direct elicitation of beliefs. The authors call this a “non-standard be-

lief mechanism” but admit that their findings could be a form of indirect reciprocity.

Their results are further extended by [Buckenmaier and Dimant \(2021\)](#) who find that willingness to cooperate in multiple social dilemma games is enhanced when subjects receive positive amounts from the initial dictator game in contrast when they receive from a computer. The authors interpret the results under the same light, that previous social experiences induce cooperation. They confine themselves to testing their hypothesis in a dictator game, trust game and public goods game. They are agnostic on the mechanisms supporting their findings. It is evident that these two studies are hinting to the effects of upstream reciprocity without conceptualizing it as such. Therefore, the present study is the first explicit study on upstream reciprocity in economics.

2.3 Experimental Design

The experimental intervention assigns 4 subjects to a reciprocity chain. The subjects are labeled as workers or helpers for convenience and they alternate in type. The first subject is a helper, the second subject is a worker, and so on. All subjects have to complete auxiliary surveys which include some control variables (demographic, positive and negative affect, social capital and economic values). The main task for workers is to perform a real effort task. The real effort task consists of an interactive graphic interface containing a matrix with zeros and ones. The matrix has a predetermined size (10×10) and the goal of the worker is to click all the zeros in the matrix. For a schematic representation of the task refer to figure [2.3.1](#). The task for the workers is to complete a predefined number of matrices to receive payment. There is no time limit for the task.

On the other hand, helpers have the option to pay a monetary amount and alleviate the burden of the workers. If a worker is helped, then the number of matrices required to receive payment is reduced. For clarity, we will use feminine pronouns for the helpers and masculine pronouns for the workers.

The game starts with a helper (subject 1), she completes the auxiliary questionnaires and then faces the decision of foregoing a portion of her endowment to reduce the workload for the next subject (subject 2, a worker), or to deny the help. Subject 2 completes the auxiliary questionnaires and then proceeds to the effort-task. After the effort task, he enters a dictator game with subject 3 (another helper) after

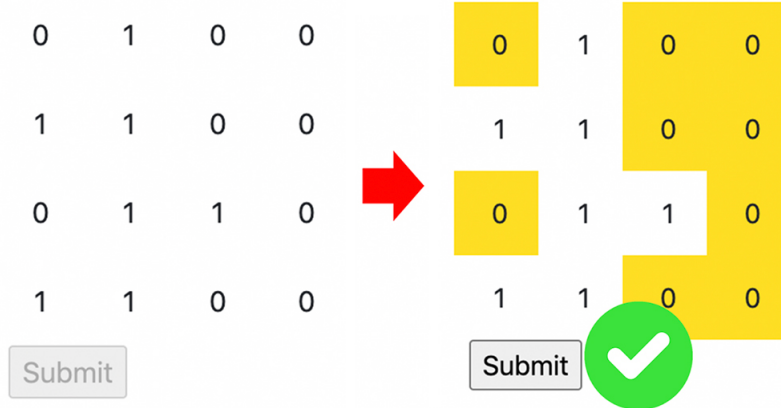


Figure 2.3.1: Schematic representation of the real effort task.

receiving an additional endowment; this creates the opportunity for a novel cooperative relationship. Finally, subject 3 can, in turn, help subject 4 (another worker). A diagram of the interaction is shown in figure 2.3.2, along with a description of the monetary payoffs to each subject. Subjects in the reciprocity chain will also be denoted by their role as Helper 1 (subject 1), Worker 1 (subject 2), Helper 2 (subject 3) and Worker 2 (subject 4).

To be more specific, Workers need to complete 15 matrices with no time limit. If Helpers decide to forego \$1 from their endowment, the number of matrices for the workers is reduced to 9. The dictator game is played with an additional endowment of \$3 and transfers are in integer amounts.

Note that the research question focuses on subject 2, since he is the player who receives a kind action and can potentially reciprocate it to a novel player. Upstream reciprocity will be measured using the amount transferred in the dictator game. Before proceeding, it is necessary to discuss some aspects of the experiment: 1) the need of a helping task, 2) the harming treatment, 3) the informational structure of the experiment and 4) the alternating tasks. A thorough description of incentives and tasks and a discussion on sample size can be found in appendices 2.A and 2.B respectively.

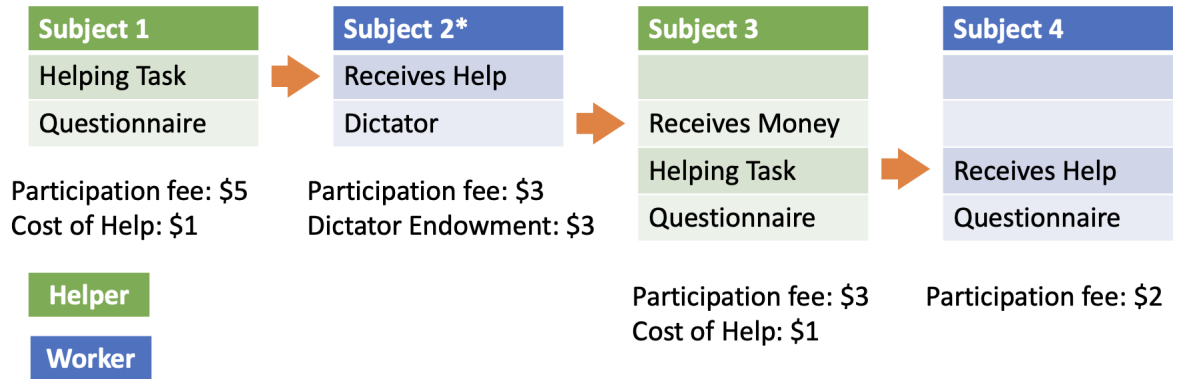


Figure 2.3.2: Upstream reciprocity chain with payment details

2.3.1 The Helping Task

Numerous experiments on altruism and empathy, specially in psychology, derive their results from a helping task (Batson et al., 1988, 1991; Schroeder et al., 1988). In those experiments, the subject faces the decision of helping someone else (often times a confederate or a fictitious subject) and how often the subject provides help is taken as a measure of prosocial behavior. In economics, the tradition is to utilize monetary incentives. Ordinary measures of prosocial behavior include using dictator games, ultimatum games or public goods games. The dictator game is often used as a measure of altruistic behavior, however such interpretation is not without challenges (Bardsley, 2008; List, 2007; Winking and Mizer, 2013; Zizzo, 2013).

However, studies of upstream reciprocity often involve a subject being both the recipient of help in a first stage and the option to provide help in a second stage. Using money as the medium of help confounds upstream reciprocity with an income effect from the help received in the first stage, or with income targeting within the experiment (Camerer et al., 1997; Cosaert et al., 2022). In other words, if a subject receives help in a first stage, it can provide more help in the second stage, not only from his intrinsic willingness to help but also because he is more able to help given that he now possesses higher monetary resources. Alternatively, if the subject has an objective monetary amount to earn from the experiment, the monetary transfer received in the first stage could cause the subject to exceed his target income and be more willing to let go of his excess income. To avoid these issues, it was necessary to implement a task where help is provided, without altering the monetary endowment

of the subject.

Hence, we opted to provide help in the form of reducing the amount of work necessary to perform a task which needs to be completed without time limit. The recipient of help can reciprocate indirectly through a monetary task; in this case, a dictator game.

2.3.2 The Harming Treatment

The experiment described previously applies to the control group. The treatment group consists of changing the helping task for a “harming task” and to modify the dictator game to allow for retaliation to an unsuspecting subject. The harming task consists of giving subjects 1 and 3 the option to increase their own payoff at the expense of increasing the number of matrices to subjects 2 and 4.

In turn, subjects 2 and 4 have the option to harm the next participant in a modified dictator game, instead of just being able to transfer to the next participant, they can also take up to a certain amount from them. This design is inspired by [List \(2007\)](#).

Details for the harm treatment are as follows: workers must complete 15 matrices as in the control group, but if they are harmed the number of matrices increases to 21. Helpers have the option to increase their payoff by \$1, if they do so the number of matrices for the matched worker will increase; the number of matrices will stay the same if they do not. Note that we have conserved the exchange of \$1 for 6 matrices. The dictator game for the workers is played with \$3, they can transfer any integer amount to the matched helper or they can take \$1. If the worker is harmed, he is notified at the end of the 21st matrix; he is informed that another subject decided to increase her payment and that now he has to complete 6 additional matrices. If the Worker is not harmed no announcement is displayed.

2.3.3 Alternating Tasks

As described in the introduction, there is a possibility that the observed upstream reciprocity is stemming from a norm. In order to mitigate the effect of “normative promiscuity” ([Haun and Tomasello, 2011](#); [Over and Carpenter, 2012](#); [Schmidt et al.,](#)

2016), we resort to changing the kind of gift received to the gift that the subject can bestow. “Normative promiscuity” is the innate tendency to infer norms where there are none, this effect is particularly pronounced in children as they are still developing their understanding of the social world. In an experimental environment, it is possible that subjects would try to infer the binding social norm. In order to mitigate this effect, the present research follows suit in the standard methodology in psychology that consists in alternating tasks (Bartlett and DeSteno 2006; Beeler-Duden and Vaish, 2020; DeSteno et al., 2010; Tsang, 2006, 2007).

2.3.4 Information Structure

Since the reciprocity chain considers a handful of subjects, it is relevant to describe what players know about one another.

- i) All subjects perform a trial of the effort task. It is important for the helpers so that they are made aware of how much they could be potentially helping. For the workers, it gives them experience with the graphic interface. Every player clicks all the zeros in just one 10×10 matrix to sense the difficulty of the task.
- ii) Each subject only knows about their own payment and costs. This is important to eliminate the possibility that subjects would attempt to equalize their payoffs motivated by inequality aversion.
- iii) Subject 1 knows about subject 2, but ignores the existence of subjects 3 and 4.
- iv) Subject 2 knows about subjects 1, 3 and 4.
- v) Subject 3 knows about subject 2 and subject 4.
- vi) Subject 4 only knows about subject 3.

Additionally, subjects that receive help are interrupted in the middle of the task. They are told that another subject paid a monetary amount to reduce the number of matrices he has to work on. Meanwhile, subjects that do not receive help are never informed that there was a subject that could have helped them. Notifying helped subjects in the middle of the task helps nullify the effect of beliefs, since the help received comes unexpected. The reason of avoiding a notification on those subjects who were not helped is to avoid feelings of resentment, and to provide a neutral comparison when their behavior is contrasted with those subjects that were helped.

In the harming treatment, subjects will be notified of receiving harm only after they have completed the task. The reasoning behind this choice is that the number of matrices is made known at the start of the effort task, which might create a reference point on the time needed to complete it. If the subject is harmed, once the task is completed, he is told that another subject increased their payment at the expense of increasing the number of matrices he has to tackle. Afterwards, the interface prompts the effort task once more. If the subject is not harmed there will be no such notification and the experiment will proceed as usual.

2.4 Estimation Procedures

Upstream reciprocity will be measured using the monetary transfer from Worker 1 to Helper 2. In order to control for other variables affecting the measure of upstream reciprocity we use a multiple regression model. Another objective of the present study consists in comparing the magnitude of upstream reciprocity under the help treatment (control group) and the harm treatment. To that objective we will use a difference-in-difference approach, using the same controls as in the multiple regression. Finally, we estimate a logit model to measure the difference in the propensity of passing the action forward, whether harm or help was received.

Since all estimation procedures use the same survey variables, it will be helpful to introduce them in the following subsection (subsection [2.4.1](#)). We will close this section with the equations to be estimated in subsection [2.4.2](#). It is important to mention that only the data from the subjects under the role of Worker 1 is considered for all estimation procedures.

2.4.1 Survey Variables

Responses from the survey will server as control variables, they are labeled *grateful*, *trust*, political spectrum index (*polspec*), a binary variable for *college* and *female*.

The variable *grateful* is obtained from the modified PANAS questionnaire. It represents the self-evaluation of gratitude in a 7 point Likert scale. The specific question is: “Indicate to what extent you feel this way right now, that is, at the present moment”. The subject then responds to 21 different affective states including “grateful”.

The subjects answer two PANAS questionnaires, the first one is asked at the start of the experiment, before the main task; the second PANAS questionnaire is asked

at the end of the experiment. Only the answer to the first questionnaire is used and hence it measures a self-evaluation of gratitude upon the commencement of the experiment. This variable is controlling for possible differential responses attributed to their base gratitude. Research in psychology has pointed out to gratitude as the main driver of upstream reciprocity (Bartlett and DeSteno, 2006; DeSteno et al., 2010; Tsang, 2006, 2007). They argue that subjects who experience higher gratitude from the favor received will also pay it forward more often and more generously.

We are agnostic on the effect of base gratitude. On one hand it could have a negative effect on the transfer, as subjects who are already feeling grateful will not interpret the help received as meriting more gratitude. On the contrary, subjects with a high base gratitude could be more prone to feeling grateful in general, in which case gratitude and the transfer will be positively correlated.

The second control variable is *trust*. Its inclusion was motivated by the social capital literature. Specifically, the subjects respond to the question “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”. The options for answer are: “Most people can be trusted” which sets the binary variable *trust* to one; alternatively selecting “Need to be very careful” sets it to zero. Trust in strangers has been shown to be the corner stone of social or civic capital. In addition, WVS question about trust in strangers correlates strongly with behavior in the trust game (Naef and Schupp, 2009), and trustworthiness (Glaeser et al., 2002; Sapienza et al., 2013). Furthermore, the WVS question about trust in strangers has been shown to be correlated with prosocial action in the public goods game (Anderson et al., 2004).

Therefore, our preliminary goal is to determine if trust, as measured through the WVS, is a predictor of upstream reciprocity. If trust is a predictor of the transfer forward then upstream reciprocity could be considered just one more among the social dilemmas that could be explained by trust. However, if trust is not a predictor of the transfer forward, then there will evidence that upstream reciprocity is a different phenomenon that enriches our current understanding of social capital. We are agnostic about the outcome of this preliminary examination.

The political spectrum index was built by summarizing the answers to the Economic values questionnaire from the World Values Survey. We use Principal Component Analysis (PCA) to build the summary. The index is built using the first principal component, the weights and further results from PCA can be consulted in

appendix [2.D](#) The index is constructed using the entire sample of Worker 1 regardless of treatment. Finally, the sign of the first principal component was switched so that “right-wing” or conservative subjects would correspond to higher values of the index (hence, moving to the right in the numerical line). Therefore, negative numbers correspond to “left-wing” or progressive subjects.

Subjects with a positive index tend to value incentives to individual effort, private ownership against public ownership and competition. The opposite is true for subjects with a negative index. Subjects with a positive index also expressed distaste for welfare and a belief that hard work leads to success. On the contrary, subjects with a negative index express affinity towards welfare and a belief that success is due to luck and connections.

Finally, we include some demographic control variables. The demographic questionnaire contains a question for education allowing for 7 categories. Due to the small sample size of this experiment, we are forced to compress the information in a binary variable for all subjects with a college degree or above, denoted *college*. Also, we asked a question on gender allowing for 3 categories: “Male”, “Female” or “Other”. For the same reason as before we were forced to reduce the information to a binary variable for the category *female*.

2.4.2 Estimation Equations

The first estimation equation regresses the transfer from Worker 1 on the help or harm received from Helper 1 and the control variables described in the previous subsection. Hence the equation is

$$transfer = \beta_0 + \beta \text{ help} + \gamma^T x + u \tag{2.1}$$

where x represents the vector of controls (*grateful*, *trust*, *polspec*, *college* and *female*), while γ represents the appropriate coefficients. Also, two versions of this equation will be considered: with and without controls for each of the treatment groups (help and harm). All throughout this section we will omit the subscript i denoting each individual observation for brevity. It is understood that all the equations refer to estimations at the individual level.

A model with interaction terms is needed in order to rightfully assess the relationship between upstream reciprocity and trust, and the corresponding relationship

with the position in the political spectrum.

$$transfer = \beta_0 + \beta \textit{help} + \lambda (\textit{trust} \times \textit{help}) + \mu (\textit{polspec} \times \textit{help}) + \gamma^T x + u \quad (2.2)$$

The coefficients of interest would be λ and μ . As before, two versions of this equation will be considered: with and without controls for each of the treatment groups for a total of four estimations. In the treatment group (harm) the variable *help* will be substituted for *harm*.

We use a logit model to estimate the difference between the propensity to pass on the helpful action and the harmful action. The dependent variable will be a binary variable that codifies when the action has been passed on. In the explanatory variables we must include a categorical variable denoting if the action received was help or harm, the marginal effect from that variable will indicate the change in probability associated with receiving harm vs receiving help. As before, we will make two versions of the model: with and without controls. The equation to be estimated is then

$$\textit{logit}(p) = \gamma^T x + \delta \textit{ur} + u \quad (2.3)$$

where p denotes the probability of passing the action forward and \textit{ur} is the indicator of which treatment the subject is. The parameter of interest is δ , which will measure the difference in propensity to pass the action forward depending on the treatment.

For the difference-in-difference (diff-in-diff) approach we need to introduce a binary variable denoting which treatment group we are considering. We call this variable \textit{ur} , as a shorthand for upstream reciprocity, and give the value of 0 for the help treatment and 1 for the harm treatment. Since we want to measure the absolute value of the difference between the two treatments we need to modify the standard diff-in-diff slightly by changing the sign of the transfer to all observations in the harm treatment. We call this new variable $\textit{transfer}^*$. Note that in doing so we are assuming that the most prevalent action in the harm treatment will be to harm forward (i.e. a negative transfer which is equivalent to taking money from the next participant).

The equation to be estimated is then

$$\textit{transfer}^* = \beta_0 + \beta \textit{help} + \gamma^T x + \delta \textit{ur} + \theta (\textit{help} \times \textit{ur}) + u \quad (2.4)$$

where \textit{ur} is an indicator variable for the treatment; $\textit{ur} = 0$ represents the control

group (help) and $ur = 1$ represents the treatment group (harm). In the regression analysis, the base category is the control group. We can now formalize the definition of $transfer^*$ since

$$transfer^* = \begin{cases} transfer & \text{if } ur = 0 \\ -transfer & \text{if } ur = 1 \end{cases}$$

Note that the coefficient for the interaction term $help \times ur$, denoted by θ , is the parameter of interest. If we denote t_0 to the transfers occurring in the control group ($ur = 0$) and t_1 to the transfers occurring in the treatment group ($ur = 1$), then

$$\theta = [E(t_0|help = 1) - E(t_0|help = 0)] - [E(t_1|help = 0) - E(t_1|help = 1)]$$

thus measuring the absolute value of the difference between the treatment effects (where treatment means receiving help or harm) in the two experimental interventions: the help treatment and the harm treatment. Note that we are working under the assumptions that $E(t_0|help = 1) - E(t_0|help = 0) \geq 0$ and $E(t_1|help = 1) - E(t_1|help = 0) \leq 0$. As before, we will estimate two versions of this model (with and without controls) in both of the experimental interventions.

It is important to mention that the dependent variable $transfer$ is originating from differing ranges in the dictator game. In the control group (help), the transfer range in the dictator game is $[0, 3]$ while in the treatment group (harm) is $[-1, 3]$. It has been shown experimentally that mean and modal responses vary with the range of the dictator game (List 2007). Precisely because of this effect, difference in means tests are not enough to quantify the differential response between treatments.

Experimental subjects are distributed in four mutually exclusive conditions: (I) Help treatment and helped, (II) Help treatment and not helped, (III) Harm treatment and harmed, and (IV) Harm treatment and not harmed. Subjects in conditions (II) and (IV) are neither helped nor harmed and they only differ in which range of the dictator game they observe. Hence, the coefficient δ is capturing the range effect exclusively. The effect between the treated subjects in both treatments is obtained by contrasting conditions (I) and (III), which is captured as $\delta + \theta$. In other words, the contrast between the treated can be decomposed as a “range effect” and a “treatment effect”.

An alternative experimental design would consist in using the dictator game with a range of $[-1, 3]$ for both treatments, help and harm. The decision to use different

ranges rests on two arguments. The first was described in the previous paragraph. By using a diff-in-diff estimation we are able to capture all the effects at play in the current interaction, including the effect for the different range. Second, an experimental intervention where a subject is helped, and then has both options to help or harm the next participant, has no precedents in the literature on upstream reciprocity and could lead to an unknown and unmeasured effect.

The reasons for such omission are clear from a psychological perspective. In a game where players interact in a forward sequence, each action conveys intent: forward play is a signaling game (McCabe et al. 2003; Smith and Wilson 2017, 2019a). A situation where a subject who has received help is confronted with the option to harm the next participant, is no longer measuring if “help has to be paid forward with help” or “harm paid with harm”. The meaning of the action is more closely interpreted as if “after receiving help you can restrain from harm”. Such change of context implies that the measurement is no longer about upstream reciprocity (paying forward help with help or harm with harm). Furthermore, giving the option to harm after receiving help could have an unknown and unmeasured effect in the subject (e.g. surprise or mistrust in the experimenter). The effect would also be unmeasurable since all the conditions would be run under the same dictator game structure, i.e. one condition would be missing making contrast impossible.

2.5 Results

A total of 88 reciprocity chains are constructed for the control group, which totals 352 subjects across the four roles. The treatment group is comprised of 86 reciprocity chains, for a total of 344 subjects. Subjects were recruited from June to October 2022 using Prolific (www.prolific.co). The graphic interface was built using oTree (Chen et al. 2016). Since the data from subjects in the role of Worker 1 and Helper 1 was critical to the experiment, their responses were filtered using attention and comprehension tests. Subjects in the role of Helper 1 faced a simple attention check while subjects in the role of Worker 1 faced the same attention check and one comprehension test related to the role of the next participant, hence testing the understanding of the reciprocity chain.

In both control and treatment, all subjects in the role of Helper 1 passed the attention check. However, two subjects in the treatment group submitted faulty data possibly due to a server error, decreasing the amount to reciprocity chains to

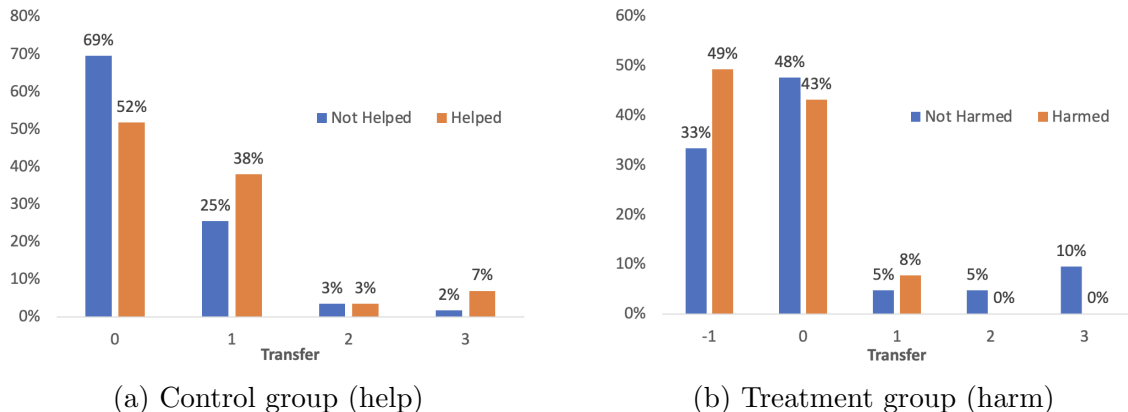


Figure 2.5.1: Distribution of transfers according to treatment.

86. Subjects in the role of Worker 1, were required to pass both the attention check and the comprehension test for their data to be considered. In all cases, subjects were allowed to finish the study and paid their participation fee and their actions carried out and paid accordingly, but were excluded from the reciprocity chain and consequently not matched with any other subject. In the control group, 90 subjects were recruited as Worker 1 and 2 failed the checks; in the treatment group 94 subjects were recruited and 8 subjects failed the checks. Further description on the demographics of the sample can be consulted in appendix [2.C](#).

2.5.1 Control Group (Help)

An exploratory inspection on the distribution of the transfers reveals a mild association between receiving help and paying forward, see figure [2.5.1a](#). Around one-third of the individuals in the role of Helper 1 decide to provide help (29/88=32.9%), from the 29 subjects receiving help around half of them decide to pay forward (14/29=48.3%). The average transfer of those who received help is 0.66, in contrast with the average transfer of 0.37 from those who did not receive help.

A simple regression yields a coefficient of 0.282 which is significant at 10%. Receiving help increases the transfer forward by almost 30 cents. See table [2.5.1](#), column (1). Once applying the appropriate controls, we see that receiving help is indeed significant at 5% and its effect is 34 cents. In a test of joint significance, the null hypothesis is rejected at 90% confidence (p-value=0.028). Significant controls

are gratitude, position in political spectrum, and college education. See table [2.5.1](#) column (2).

Gratitude emerges as one relevant control, specially in light of the results from the psychology literature. An increase in one point in self-assessed gratitude upon starting the experiment reduces the transfer forward by 12 cents. Initially, this might seem counter-intuitive as more grateful subjects should transfer more, however the mechanism is more subtle. As subjects assess their gratitude higher, they also have a lower evaluation of the help received, hence reducing the transfer forward.

To verify this claim consider a regression that explains the gratitude differential through the initial measure of gratitude. As described in section [2.4.1](#) subjects answer two affect questionnaires, we use the difference between the measures as dependent variable; the initial measure is the independent variable. All units are in Likert points. The regression yields a coefficient of -0.187 with a p-value of 0.027. Full regression results can be found in appendix [2.E.1](#). This means that for every additional point on base gratitude, the differential measure drops by almost one fifth of a point. Considering that the average measure of the gratitude differential is 0.32 and that the standard error is 0.11, the estimated magnitude is relevant. Hence, the higher the gratitude of the subject upon entering the experiment, he or she is less susceptible to feel gratitude from the intervention.

Note that other significant controls are the political spectrum index (*polspec*), which means that for every point that the subject is leaning towards the conservativeness, the amount paid forward decreases by around 15 cents. College education is playing a significant role by increasing the transfer by approximately 30 cents, as high as the effect of having received help.

2.5.2 Treatment Group (Harm)

In the treatment group there is a much stronger relationship between the intervention and the forward transfer. Interestingly enough, around three quarters of the subjects in the role of Helper 1 decide to harm the next participant ($65/86=75.6\%$). In very colloquial terms, evil gets a head start. Going forward, around one half of those receiving harm decides to pass the harm forward ($32/65=49\%$). The average transfer for those who did not receive harm was 0.10 while it was further reduced to -0.42 for those who received harm. See figure [2.5.1b](#)

	Help (control)		Harm (treatment)	
	(1)	(2)	(3)	(4)
helped (or harmed)	0.282* (0.163)	0.344** (0.154)	-0.511** (0.204)	-0.448** (0.212)
grateful		-0.120*** (0.043)		-0.076 (0.047)
trust		0.232 (0.145)		-0.230 (0.194)
polspec		-0.161*** (0.046)		-0.003 (0.053)
college		0.316** (0.143)		0.265 (0.184)
female		-0.103 (0.149)		0.135 (0.184)
Constant	0.373*** (0.093)	0.611** (0.240)	0.095 (0.178)	0.256 (0.282)
Observations	88	88	86	86
R-sq	0.0338	0.2297	0.0693	0.1519
F-stat	3.0046	4.0261	6.2529	2.3587
P-value	0.0866	0.0014	0.0143	0.0379

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5.1: Estimation results summary.

The simple regression describes a decrease of 51 cents in the average transfer after receiving harm, and with controls the effect dampens to a 45 cent decrease. Both coefficients are significant at 5%. See table 2.5.1 column (3). In the propagation of harm, we start observing a certain asymmetry since none of the controls appear to be significant. The effect of political spectrum index is negligible, which indicates that political spectrum is a significant predictor of positive upstream reciprocity but it is not a predictor of negative upstream reciprocity. The effect of education is again positive and similar in magnitude as in the control group, however the coefficient is not significant.

2.5.3 Interaction Terms

The results from the estimation with interaction terms can be found in table 2.5.2. For the control group (help), the interaction terms with trust are significant. Note that the magnitude of the coefficient (around -55 to -65 cents) is similar in magnitude to those of the previous interaction, i.e. having received help (around 60 to 70 cents); also note that they have opposite signs. This reveals an interesting mechanism of upstream reciprocity. Subjects who trust strangers will transfer around 40 cents more than untrusting subjects, and the effect of the previous interaction will be negligible, considering that the effect of having received help will cancel out with the effect of trusting others. In contrast, subjects who do not trust in strangers will see their transfers increased only when receiving help in the previous interaction.

This finding uncovers a mechanism through which upstream reciprocity enhances cooperation within the group. The evidence from numerous experiments in economics consistently discovers two types of subjects: the selfish-type and the pro-social type². The motives behind the observed pro-sociality are diverse and sometimes vary with context. Theoretical literature in social biology posited that upstream reciprocity enhances cooperation within the group, leading to higher group fitness (Nowak and Roch, 2007), but the micro-structural mechanisms were still unexplained. As described in the introduction, trust has emerged consistently as a correlate of pro-sociality in various social dilemma games (Glaeser et al., 2002; Anderson et al., 2004; Naef and Schupp, 2009; Sapienza et al., 2013). Hence, results in this study suggest that upstream reciprocity changes non-cooperators into cooperators, where the pre-

²The literature has provided numerous examples, some of them summarized in (Fehr and Gintis, 2007) and more recently in (Cooper and Kagel, 2016).

vious social interaction supplements the lack of trust. The net effect is an increased level of cooperation within the group.

Finally, the interaction effect between the position in the political spectrum and the previous interaction is not significant. This means that subjects with preferences for progressive policies have a higher transfer forward on average, and the effect is independent of the previous interaction. As in the previous case, we are not able to uncover any mechanism for negative upstream reciprocity.

2.5.4 Comparison across treatments

The most simple comparison is to compute the survival of help or harm along the chain. In the control group, a kind act is started in 32.9% of the cases (initiation rate) and then is continued forward in 48.3% of the cases (propagation rate). In total, 15.9% of the chains maintain positive reciprocity up to stage 2 (survival rate). For comparison, in the treatment group, subjects harm the following player in 75.6% of the cases, the unkind action is passed forward in 49.2% of the cases. In total, 37.2% of the chains maintain negative reciprocity up to stage 2. Results are summarized in table [2.5.3](#).

If we focus on subjects in the role of Worker 1 and compare the rates at which the previous action is passed forward, whether if it is harm or help, a difference in means test fails to reject the null hypothesis that the rates are different (p-value=0.932). Hence, harmful or helpful actions are passed forward at approximately the same rate.

To further formalize this comparison and reinforce the results, we run a logit regression where the dependent variable is the probability of passing the action forward regardless if it is kind or unkind. We run two models, with and without controls. The main explanatory variable is if the subject belongs to the help treatment or the harm treatment. The base category is belonging to the help treatment. Results show that the regressions are overall not significant and therefore no significance difference can be established between passing harm or help forward. Results are summarized in table [2.5.4](#) where average marginal effects are reported. Note that trust is associated with a marginal increase in the propagation rate, and so does a preference for progressive policies.

The next step consists in a comparison between the coefficients from the estima-

	Help (control)		Harm (treatment)	
	(1)	(2)	(3)	(4)
trust	0.487*** (0.176)	0.423** (0.178)	-0.363 (0.513)	-0.385 (0.543)
helped (or harmed)	0.706*** (0.212)	0.622*** (0.216)	-0.439* (0.242)	-0.448* (0.247)
trust × helped	-0.657** (0.307)	-0.556* (0.310)	0.188 (0.555)	0.179 (0.575)
polspec	-0.139** (0.054)	-0.143*** (0.054)	-0.065 (0.110)	-0.075 (0.109)
polspec × helped	-0.051 (0.089)	-0.057 (0.088)	0.064 (0.126)	0.096 (0.127)
grateful	-0.129*** (0.042)	-0.128*** (0.043)	-0.090* (0.048)	-0.081 (0.050)
college		0.276* (0.143)		0.281 (0.188)
female		-0.075 (0.149)		0.144 (0.195)
Observations	88	88	86	86
R-sq	0.2337	0.2686	0.1251	0.1594
Adj. R-sq	0.1770	0.1946	0.0587	0.0721
F-stat	4.1175	3.6272	1.8828	1.8258
P-value	0.0012	0.0012	0.0941	0.0849

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5.2: Interaction terms estimation summary.

	Initiation	Propagation	Survival
Help	33.0%	48.3%	15.9%
Harm	75.6%	49.2%	37.2%
Difference (p.p.)	42.62***	0.01	21.30***
p-value	0.000	0.932	0.002

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5.3: Rates for different variables of interest in the reciprocity chain

	(1)	(2)
help to harm	0.373 (0.310)	0.518 (0.325)
grateful		-0.0728 (0.0877)
trust		0.726** (0.335)
polspec		-0.170* (0.0997)
college		-0.0144 (0.326)
female		-0.282 (0.329)
N	174	174
Pseudo R^2	0.00619	0.0411
LR χ^2	1.456	9.659
p-value	0.228	0.140

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5.4: Average marginal effects from logit regressions to explain the propagation rate.

Significance codes: * 10%, ** 5%, *** 1%.

	Help (control)	Harm (treatment)	Simple Difference	Diff-in-Diff
No Controls	0.282* (0.163)	- 0.511** (0.20)	0.229 0.482	0.228 0.380
With Controls	0.344* (0.154)	- 0.450** (0.27)	0.240 0.558	0.189 0.474

Standard errors in parentheses and p-values are without parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5.5: Comparison table.

tions of equation [2.1](#). In other words, we will compare the coefficients of regression (1) vs (3), and (2) vs (4), as described in table [2.5.1](#). The objective is to measure the difference in the magnitude from the individual responses of having received help vs having received harm. We make this comparison by means of a chi-squared test between the coefficients of the models with and without controls. Results are summarized in table [2.5.5](#), under the column labeled “Simple Difference”. In the models without controls, the difference in transfers between help and harm is approximately 23 cents. This means that receiving help increases the forward transfer by 28 cents but receiving harm decreases it by 51 cents. We are interested if the differences are equal but with opposite signs. The chi-square test fails to reject the null hypothesis with a p-value of 0.482. The same analysis is repeated for the model with controls. Again, the test fails to reject the null hypothesis with a p-value of 0.558. Note that the standard errors are too large, which might be the cause of the failure to reject the null hypothesis.

The final step consists in the difference-in-difference (diff-in-diff) estimation (see equation [2.4](#)). Recall we have two groups: control (help) and treatment (harm). In each group, the treated individuals correspond to those receiving help or harm respectively. The diff-in-diff estimator allows to compare the treatment effects (receiving help against not receiving help, and receiving harm against not receiving harm), between the treatment and control group. Recall that to account for the opposite sign of the effects, the dependent variable was made negative for observations in the harm treatment. Information on the full diff-in-diff regression can be consulted in section [2.E.2](#)

In table [2.5.5](#), the column labelled “Diff-in-Diff” summarizes the result from the difference-in-difference estimation to compare the absolute value of the forward trans-

fer between each treatment; which corresponds to the coefficient θ in equation [2.4](#). When comparing models without control variables, treatment effects are approximately 23 cents apart, this magnitude cannot be distinguished from zero (p-value: 0.380). When control variables are included, treatment effect is of similar size around 19 cents and not statistically different from zero (p-value: 0.474). Most likely, the results are caused by the large standard errors. In the estimate without controls the standard error is 0.259, and 0.263 for the estimate with controls. Hence, results from this test are inconclusive. Future measurements of upstream reciprocity should focus on obtaining more precise estimates to allow for comparison. Another avenue of research consists in a proper understanding of the quantitative relationship between the magnitude of the gift received and the magnitude of the gift paid forward.

2.6 Conclusion and Discussion

Upstream reciprocity is a key element in the evolution of cooperative behavior and direct reciprocation ([Nowak and Roch, 2007](#)). Previous research in psychology had shown its prevalence and discussed its mechanisms, with gratitude emerging as the main cause. In economics, recent research had shown that previous social interaction can have an effect in the propensity to cooperate in social dilemma games, pointing to upstream reciprocity as the likely explanation. The present research is focused on its relationship with trust and social capital.

The main conclusion is that upstream reciprocity enriches our current understanding of social capital. We find that upstream reciprocity supplements trust: individuals who exhibit a lack of trust behave as prosocially as trusting individuals but only after a previous positive interaction. It is yet to be determined if there are differences in the application of upstream reciprocity across cultures. Such avenue of research would further uncover how upstream reciprocity, as an aspect of social capital, furthers societies in their design of institutions and economic outcomes. Opinions in the political spectrum are also an important predictor of positive upstream reciprocity exclusively, subjects with “progressive” preferences have a higher transfer forward on average. Gratitude appears as an important predictor of positive upstream reciprocity, in line with the results from psychology. Although our conclusions are drawn from a different methodology.

When comparing positive and negative upstream reciprocity we find that kind or unkind actions are passed forward at approximately equal rates (extensive mar-

gin), and at approximately equal magnitudes but in opposite directions (intensive margin). Extensive margin results are supported through simple difference in means test and logit regression. Intensive margin results are obtained from a difference-in-difference regression across treatments, considering control variables as well as difference in means tests. Results in the intensive margin are limited due to large standard errors. Further research can focus on refining the measurements on the extensive margin, not only to decrease the size of the standard errors but to uncover the relationship between the magnitude of the gift passed forward and its action space, and with the magnitude of the gift received.

It is also important to mention that no relevant predictors to negative upstream reciprocity were found, which highlights asymmetric motivations to propagating kindness or unkindness. The only relevant predictor of negative upstream reciprocity is having received an unkind act in the previous stage. To understand the motivations behind negative indirect reciprocation, more accurate instruments are needed whether in the form of psychological questionnaires or incentivized interventions.

The mathematical modeling of upstream reciprocity poses certain challenges. If choosing a utilitarian model with an altruistic component, one would have to consider that the altruistic parameter can be influenced by external forces, in this case the reception of the initial act of kindness or unkindness. Alternatively, as previous research suggests, previous “social experiences” may warp beliefs about other’s propensity to cooperate as found by [Buckenmaier and Dimant \(2021\)](#); [Schwerter and Zimmermann \(2020\)](#). Therefore, a belief approach might be more suitable for modeling upstream reciprocity with certain reservations. The main challenge from a utilitarian perspective is that upstream reciprocity would need to be modeled as an increased utility derived from bestowing an act of kindness to a third party who is independent of the previous interaction, or as an increased utility from harming a third party; considering that there are no prospects of future interactions with the third party. Such increase in utility would have to be independent of payoffs and derived purely from executing an action rather than its outcome, which is contradictory to the spirit of utility modeling.

If one would like to opt for framing upstream reciprocity under Kantian equilibrium ([Roemer, 2010](#)), one faces the contradiction that continued cooperation could easily be modeled as such while continued harm is contrary to the categorical imperative: “act only in accordance with that maxim through which you can at the same time will that it become a universal law” ([Kant, 2002](#)). On the other hand,

the prevalence of the propagation of harm shown in this study, even at a higher absolute rate than the propagation of kindness, can be easily understood under a purely selfish paradigm, but it leaves the propagation of kindness unresolved.

In opposition to such approaches, the present research further supports the role of gratitude as understood by Adam Smith in his moral sentiments (Smith, 1976) and it is more akin to the unorthodox approach of Vernon Smith and Bart Wilson who stress the mechanism of “beneficence leads to gratitude, which leads to reward” and “harm leads to resentment, which leads to punishment” (Smith and Wilson, 2017, 2019a). Furthermore, it shows that the recipient of the reward or punishment could be anyone in the society.

Appendix

2.A Incentives and Tasks

It is necessary to explain the incentives and costs for each subject, considering there are four subjects with varying roles. Also, it is important to mention that for the sake of running this experiment at a reasonable costs, the questionnaires needed for each role are different, leading to different times required for every role to complete their task. A complete list of tasks will be provided in this section as well as the monetary incentives and costs for every role in both treatment and control groups. A full reproduction of the questionnaires employed can be consulted in the appendix [2.F](#).

The following questionnaires are employed:

a) **PANAS - Positive and Negative Affect Schedule**

Self-report questionnaire consisting of 10 items related to positive and negative affect. Two additional items evaluating for gratitude and resentment were needed. They were worded as 'grateful' and 'resentful'. The instantaneous version was used, i.e. subjects are asked to rate their present emotional state ("Indicate to what extent you feel this way right now, that is, at the present moment") (Watson et al., 1988).

PANAS is used to determine the base gratitude and resentment that the subject is bringing to the experiment. Controlling for this variable will allow to measure the effect of the treatment. Also, subjects labeled as Worker 1 need to complete a second PANAS to measure if there are any affective states that are being influenced by the experiment.

b) **Social Capital**

Questionnaires on Social Capital, Trust and Organizational Membership from the

World Value Survey is applied. The questionnaire used in this study excludes the questions related to organizational membership. We are interested in measuring if upstream reciprocity is associated to self-reported measures of interpersonal trust exclusively. The questionnaire includes one key question asking “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”, with a yes or no answer options. This question is followed by a ranking of trust among several institutions. Ranking is answered in a 4 point Likert and the institutions are: family, neighborhood, people you meet for the first time, people of other nationality and people of other religion. We use version 2017-2021 Wave 7 (Haerpfer et al., 2020).

c) **Economic Values**

The questionnaire on Economic Values from the World Values Survey is applied. In this case, questions are related to the opinions of the subject around the topics of equality vs incentives for individual effort, private vs public ownership, government welfare, competition and beliefs about hard work leading to success. Questions are answered on a 10 point Likert scale, where each end of the scale corresponds to an opposing opinion. For example, in the question related to equality vs incentives: 1 corresponds to maximal adherence to “Incomes should be made more equal”, while a 10 corresponds to maximal adherence to “There should be greater incentives for individual effort”.

d) **Demographic questionnaire**

Includes questions for age, income and gender.

In addition to the questionnaires, there are four tasks:

a) **Real Effort Task**

The subject faces a sequence of square arrangements containing only zeros and ones. The task consists in clicking all the zeros. The subject faces one arrangement at a time and it is not allowed to advance to the next arrangement until all the zeros (and only the zeros) have been clicked. The subject needs to complete 15 square arrangements for the task to be considered complete.

b) **Real Effort Task Trial**

The subject faces only one rectangular arrangement like the ones described in the real effort task. The trial comes with specific instructions. The objective of the task is to familiarize the workers with the interface and to give the helpers a sense of how much they will be helping (or harming) the workers. Each arrangement is of dimension 10×10 .

	Helper 1	Worker 1	Helper 2	Worker 2
Task Sequence	PANAS	PANAS	Economic Values	Economic Values
	Social Capital	Social Capital	Effort Task Trial	Effort Task Trial
	Economic Values	Economic Values	Helping Task	Effort Task
	Effort Task Trial	Effort Task Trial		
	Helping Task	Effort Task		
	Demographic	PANAS		
		Demographic		
Participation Fee	\$3	\$3	\$3	\$2
Cost of Help/Harm	\$1	x	\$1	x
Dictator Endowment	x	\$3	x	x

Table 2.A.1: Summary of task sequence and monetary incentives for each role.

c) **Helping/Harming Task**

The subject faces the option to pay a portion of her participation fee and reduce the number of the square arrangements the next player is facing. The cost of helping the next subject is \$1 and it reduces the number of square arrangements from 15 to 9.

In the harming task, the subject faces the option to increase his or her own payoff by \$1. If they decide to do so, the number of square arrangements of the next participant are increased from 15 to 21.

d) **Dictator Game**

The subject receives a monetary endowment of \$3 and decides how much to keep, the rest will be transferred to the next player. In the treatment group, where subjects have the ability to harm each other, the subject has the additional option to take up to \$1 from the next participant

The order of the tasks for each subject as well as their monetary incentives is summarized in table [2.A.1](#)

2.B Sample Size Determination

Computations of sample size were based on [Beeler-Duden and Vaish \(2020\)](#) since it employs a very similar design and provides detailed information on the means and

standard deviations on the second subject, which is the main subject of study. In their study, upon receiving help, subject 2 sends 1.6 experimental currency units to subject 3 with a standard deviation of 1.5. When subject 2 does not receive help, 0.45 units are sent with a standard deviation of 0.94. Determination of sample size employs Wilcoxon Mann-Witney test for two groups (Cohen's $d=0.926$) with an allocation ratio of 0.17.

To estimate the allocation ratio, we use a stressed scenario where most subjects do not provide help, when help is costly and there is a justification to not provide it. Utilizing a monetary cost to reduce the burden on another experimental subject has no precedents in psychology and economics, at least to the knowledge of the author. To provide an estimate we use [Batson et al. \(1988\)](#) who found that only 15% of subjects would provide costly help, this yields a ratio of 0.17.

Most experiments in upstream reciprocity use a fictitious subject 1, which allows for an external manipulation on how many subjects 2 will receive help. Since deception is anathema in Economic experiments, we must adjust the ratio considering that subjects 1 are free to choose top help or not.

Using a confidence level of 5% and power of 80% yields a sample size of 64 for the control group. Using a power of 90% yields a sample size of 86. Note that this is the number of reciprocity chains, hence the number must be multiplied by 4 to account for all the subjects involved in the chain. This yields a total number of subjects to 256 and 344 for 80% and 90% power respectively.

Finally, the treatment design is partially based on the experiment Berk15 in [Charness and Rabin \(2002\)](#). More specifically, subject 2 will perform a similar dictator game as the one Berk15, where it was found that approximately 27% take the self-centered action. Notice that the proportion is dramatically different from those reported by [Batson et al. \(1988\)](#). This could allow for a reduced sample; however, the size will be kept the same as in the control group for simplicity, comparability and to reduce the risk of an underpowered experiment. For that reason, sample size will remain the same as in the control group at a total number of subjects to 256 and 344 for 80% and 90% power respectively.

The final experiment was run under 90% power. All subjects were recruited by using Prolific, with the additional condition of a balanced sample (i.e. the sample is balanced between male and female participants). Only subjects with verified Prolific

	Help (control)	Harm (treatment)	Total
Female	44	43	87
Male	43	42	85
Other	1	1	2
Total	88	86	174

Table 2.C.1: Gender distribution by treatment group. Subjects in the role of Worker 1 exclusively.

accounts with residency in the United States and proficiency in the English language are allowed to participate. Subjects that failed attention checks or comprehension tests were allowed to finish the experiment but their data was excluded from the analysis.

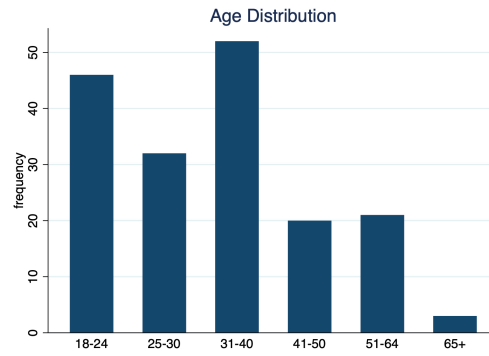
2.C Demographics from Sample

Sample was drawn from Prolific, a crowdsourcing website focused on survey research. Sample is not representative of the U.S. population, however it is still drawn from a general audience. Several demographic variables are in display in table [2.C.1](#) and figure [2.C.1](#) to test for the external validity of the results. The sample was drawn using the automatic gender balance feature.

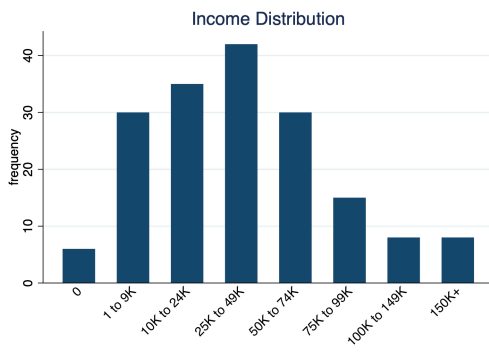
2.D Principal Component Analysis

The political spectrum index (*polspec*) was constructed from a principal component analysis (PCA) from the responses to the Economic Values survey questions, and extracting the first component. Last, *polspec* is the negative of the first component so that subjects on the “right-wing” of the spectrum would lie in the positive numbers, and hence to the right on the number line. The results from principal components are laid out in tables [2.D.1](#) and [2.D.2](#).

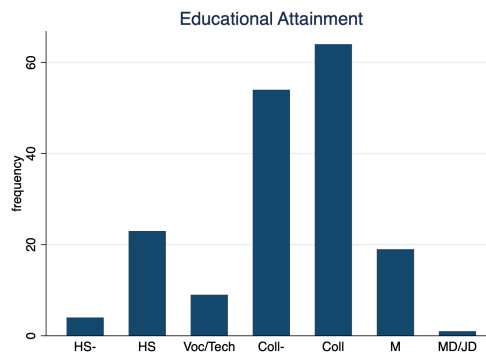
Finally, to verify the results from the PCA analysis we used a K-nearest neighbors to classify all the subjects into two groups, using the raw data from the Economic Values survey questions. Both methods agree on their classification in 90% of the cases, which represents a disagreement only on 17 subjects. This gives us confidence that the political spectrum represents our notions of left and right opinions. The



(a) Age Distribution



(b) Income Distribution



(c) Educational Attainment

Figure 2.C.1: Demographic variables of subjects in the role of Worker 1

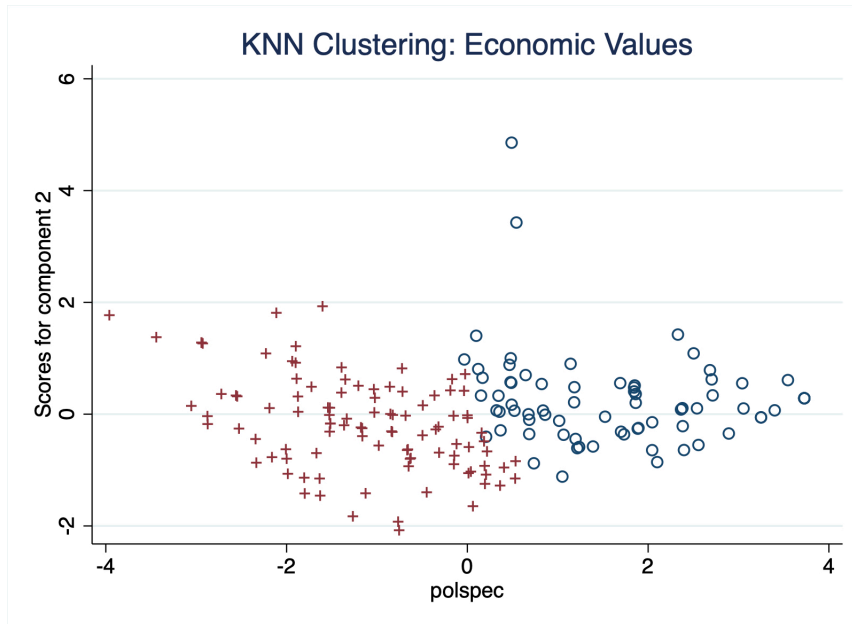


Figure 2.D.1: Overlay of clustering and principal component analysis. Second principal component is shown for easier visualization.

classification methods can be contrasted in figure [2.D.1](#)

2.E Additional Estimation Results

This section contains additional regression results. Subsection [2.E.1](#) contains a brief description and results of the gratitude mechanism: subjects with a high self-reported gratitude at the beginning of the experiment experience less gratitude from the inter-

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.93452	2.18189	0.5869	0.5869
Comp2	0.752625	0.185969	0.1505	0.7374
Comp3	0.566656	0.109547	0.1133	0.8508
Comp4	0.457109	0.168019	0.0914	0.9422
Comp5	0.28909	.	0.0578	1

Table 2.D.1: Eigenvalue decomposition from Principal Component Analysis

Variable	Comp1	Comp2	Comp3	Comp4	Comp5
equality	-0.4964	0.3029	0.1801	0.2097	0.7651
ownership	0.4265	0.4383	0.5851	-0.5215	0.1085
welfare	-0.4712	0.3562	0.4343	0.2716	-0.6234
competition	0.3902	0.6986	-0.4435	0.4026	-0.0294
hardwork	0.4443	-0.3181	0.4898	0.6695	0.1154

Table 2.D.2: Scoring coefficients from Principal Component Analysis

vention. In subsection [2.E.2](#) we report the full results from the difference-in-difference estimation described in equation [2.4](#).

2.E.1 Gratitude Mechanism

The gratitude mechanism is a result relevant to the literature in psychology. It is indeed one mechanism operating in upstream reciprocity, however its effect is counter-intuitive. The more gratitude the subject experiences upon the commencement of the experiment, the less gratitude will be experienced from the intervention.

To verify this claim we rely on the two self-reports of gratitude in 10-pt Likert scale. The variable *grateful*, used throughout this study, will be complemented with the variable *grateful_end*, which denotes the self-reported gratitude at the end of the experiment. The equation to be estimated is:

$$grateful_end - grateful = \xi_0 + \xi grateful + w \quad (2.5)$$

To avoid confusion we will refer to *grateful* as "base gratitude" and *grateful_end - grateful* as the "gratitude differential". Results are found in table [2.E.1](#)

2.E.2 Difference-in-Difference

This subsection contains the complete estimation results from the difference-in-difference analysis. Results are found in table [2.E.2](#) column (1) contains the results without controls and column (2) contains the results with controls.

The estimation equation [\(2.4\)](#) is reproduced below for reference:

$$transfer^* = \beta_0 + \beta help + \gamma^T x + \delta ur + \theta (help \times ur) + u$$

	(1) Gratitude Differential
Initial Gratitude	-0.187** (0.0828)
Intercept	0.993*** (0.373)
N	88
R^2	0.0558
Adj. R^2	0.0449
F-stat	5.086
p-value	0.0267

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.E.1: Results for the gratitude mechanism (equation 2.5)

2.F Transcripts of Experimental Procedure

2.F.1 Consent Form

[Consent forms vary according to role and payment.]

2.F.2 PANAS

Below you will find a scale and a series of words that describe different feelings and emotions. Read each item and then mark the appropriate answer using the radial buttons.

Indicate to what extent you feel this way right now, that is, at the present moment:

[The subject answers the following items in a 7 point Likert scale: not at all / very slightly / a little / moderately / quite a bit / very / extremely. Items are randomized for every subject]:

1. Interested
2. Distressed
3. Excited
4. Upset
5. Strong
6. Guilty
7. Scared
8. Hostile
9. Enthusiastic
10. Proud
11. Irritable
12. Alert
13. Ashamed
14. Inspired
15. Nervous
16. Determined
17. Attentive
18. Jittery
19. Active
20. Afraid
21. Grateful

	(1)	(2)
helped (or harmed)	0.282 (0.174)	0.298* (0.177)
ur	-0.468** (0.195)	-0.412** (0.201)
helped x ur	0.228 (0.259)	0.189 (0.263)
grateful		0.003 (0.032)
trust		0.190 (0.124)
college		0.003 (0.121)
female		-0.055 (0.121)
Observations	174	174
R-sq	0.0657	0.0903
Adj. R-sq	0.0492	0.0462
F-stat	3.9820	2.0480
P-value	0.0090	0.0438

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ur represents the indicator variable of treatment:

$ur = 1$ for control group (help)

$ur = 2$ for treatment group (harm)

Base category is $ur = 1$.

Table 2.E.2: Results for diff-in-diff estimation (equation 2.4)

2.F.3 Social Capital Questionnaire

Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?

Binary Answer:

- a) Most people can be trusted
- b) Need to be very careful

You will be presented with a list of various groups. For each one, could you tell whether you trust people from this group completely, somewhat, not very much or not at all?

[Answer in 4-point Likert scale: Trust completely / Trust somewhat / Do not trust very much / Do not trust at all]:

- i. Your Family
- ii. Your Neighborhood
- iii. People you know personally
- iv. People you meet for the first time
- v. People of another religion
- vi. People of another nationality

2.F.4 Economic Values Questionnaire

Now we would like to know your views on various issues. How would you place your views on this scale?

1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between.

[Answers in 10 point Likert scale]:

Incomes should be made more equal	There should be greater incentives for individual effort
Private ownership of business and industry should be increased	Government ownership of business and industry should be increased
Government should take more responsibility to ensure that everyone is provided for	People should take more responsibility to provide for themselves
Competition is good	Competition is harmful
In the long run, hard work usually brings a better life	Hard work doesn't generally bring success—it's more a matter of luck and connections.

2.F.5 Effort Task

In the following pages you will be asked to perform the point-and-click task. You will be shown square arrangements of cells containing zeros and ones. Each arrangement will be called a matrix (plural: matrices). An example of such arrangement is shown below.

[Image of a small matrix shown]

Every time you see a matrix, your task consists of clicking all the cells containing a zero. Once you click on a cell, it's status changes to "activated" and will change color. After selecting all the zeros you will be allowed to advance to the next matrix.

Your task consists of selecting all the zeros in a total of 15 matrices.

Each matrix is an arrangement of 10 x 10 cells.

You will not be able to advance if you have activated a cell that does not contain a zero or if you have not yet activated a cell containing a zero.

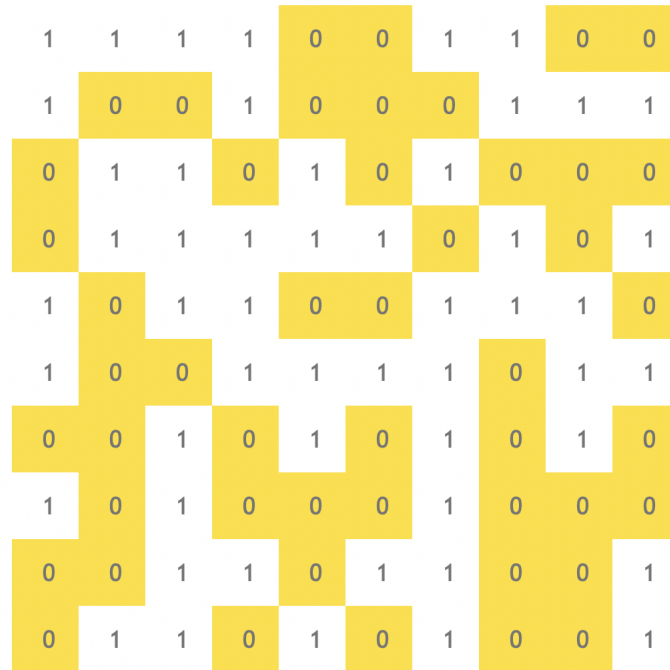
[A diagram on how the interface works is shown]

In the next page, you will practice the point-and-click task with only one matrix.

The objective is to familiarize yourself with the interface.

MATRIX 5 OF 15

Select all the zeros (and only the zeros):



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

Next

Figure 2.F.1: Sample of effort task graphic interface

[One matrix is shown and the subject has to click all the zeros]

Now you will proceed with the point-and-click task.

[The effort task begins. Refer to figure 2.F.1 for sample interface.]

[If subject receives help, he is shown a notice like in figure 2.F.2. The effort task concludes afterwards. If subject is harmed, upon the completion of the 15 matrices, a notice like in figure Y is shown. The subject is then prompted back to the effort task with the counter set at 15 out of 21. Notice that both are followed by an attention check.]

ATTENTION!

Another Prolific participant who previously answered another survey gave up a portion of their payment to reduce the number of matrices you were assigned.

This means that your point-and-click task is over.

Please answer the following question as an attention check.

How do you like the weather today? Irrespective of how do you like the weather, please select 'Average'. This is an attention check.

- Excellent
- Good
- Average
- Poor

Next

Figure 2.F.2: Notice after receiving help.

ATTENTION!

Another Prolific participant who previously answered another survey decided to increase her own payment at the expense of you having to perform the point-and-click task for 6 additional matrices.

You will be redirected to the point-and-click task.

Please answer the following question as an attention check.

How do you like the weather today? Irrespective of how do you like the weather, please select 'Average'. This is an attention check.

- Excellent
- Good
- Average
- Poor

Next

Figure 2.F.3: Notice after receiving harm.

2.F.6 Demographic Questionnaire

Please indicate the highest level of education completed.

- a) Less than high school
- b) High school or equivalent
- c) Vocational/Technical School (2 years)
- d) Some college
- e) College graduate (4 years)
- f) Master's degree (MS, MA, etc)
- g) Doctoral degree (PhD)
- h) Professional degree (MD, JD, etc)

What is your gender?

- a) Male
- b) Female
- c) Other

How old are you?

- a) Under 18
- b) 18 - 24
- c) 25 - 30
- d) 31 - 40
- e) 41 - 50
- f) 51 - 64
- g) 65 or older

Which of these describes your personal income last year?

- a) \$0
- b) \$1 to \$9 999
- c) \$10 000 to \$24 999
- d) \$25 000 to 49 999
- e) \$50 000 to 74 999
- f) \$75 000 to 99 999
- g) \$100 000 to 149 999
- h) \$150 000 and greater

2.F.7 Dictator Game

Recall that you receive a participation fee of \$3 for completing this study.

Since you have reached this point in the study, you earn an additional \$3.

[page break]

As part of this study, the experimenters have recruited other participants to answer a different questionnaire. You will be matched with one of them after you complete this study.

For the sake of clarity, we will refer to that player as “the next participant”.

[page break]

The next participant has the option to give out a portion of her payment to make the clicking task less burdensome to another participant in your role.

If the next participant gives out a portion of his/her payment, the number of rounds for another participant will be reduced to 9, *just as it happened to you*. [The emphasized sentence is removed if the subject did not receive help. Emphasis is added in

Before proceeding, answer this question to the best of your knowledge. This is a comprehension check.

Complete the sentence with the correct statement.

The next participant...

- Has to work through a point-and-click task, just like yourself.
- Has the option to pay a portion of his/her payment to make the clicking task less burdensome to another participant in your role
- Has the option to send a monetary amount to another participant in your role.
- Has paid a portion of his/her endowment to make the point-and-click task less burdensome to you.

[Click here to expand/collapse the instructions](#)

Next

Figure 2.F.4: Comprehension test

the graphic interface to call the attention of the subject].

Your next task is to decide how much of the additional \$3 to transfer to the next participant.

[Comprehension test is shown afterwards in a new page. See figure 2.F.4. Feedback from their answer in the comprehension test is given immediately afterwards]

[Followed by a new page with the dictator game]

In this page you will choose the transfer to the next participant.

Your total income from this survey will be equal to \$3 plus the amount of money you keep for yourself.

Choose your transfer to the next participant, you can transfer from \$0 to \$3 in dollar amounts (no decimals):

[Text entry box]

2.F.8 Dictator Game: Treatment Group

Recall that you receive a participation fee of \$3 for completing this study.

Since you have reached this point in the study, you earn an additional \$3.

[page break]

As part of this study, the experimenters have recruited other participants to answer a different questionnaire. You will be matched with one of them after you complete this study.

For the sake of clarity, we will refer to that player as “the next participant”.

[page break]

The next participant has the option to increase their own payment at the expense of making the point-and-click task more burdensome to someone else in your role.

If the next participant decides to increase his/her payment, the number of matrices for someone else in your role will be increased to 21, *just as it happened to you*. [The emphasized sentence is removed if the subject did not receive harm. Emphasis is added in the graphic interface to call the attention of the subject].

In your next task, you have the chance to influence the monetary payoff of the next participant.

Your next task is to decide how much of the additional \$3 to *transfer* to the next participant, or if you decide to do so you can also *take \$1* from the next participant. You also have the option to neither take nor transfer. [Emphasis is added in the graphic interface to call the attention of the subject]

[Comprehension test is shown afterwards in a new page. See figure [2.F.5](#). Feedback from their answer in the comprehension test is given immediately afterwards]

[A new page follows with the dictator game.]

In this page you affect your monetary payoff and that of the next participant.

Before proceeding, answer this question to the best of your knowledge. This is a comprehension check.

Complete the sentence with the correct statement.

The next participant...

- Has sent a monetary transfer to you.
- Has the option to send a monetary amount to another participant in your role.
- Has the option to increase his/her payment which would increase the number of matrices to someone else in your role.
- Has paid a portion of his/her endowment to make the point-and-click task less burdensome to you.

[Click here to expand/collapse the instructions](#)

Next

Figure 2.F.5: Comprehension test for the treatment group: harm

Recall that your participation fee is \$3 and that you have received an additional \$3, which can be used to make a transfer to the next participant.

If you choose to take \$1, we will decrease the payment of the next participant by \$1 and increase yours by \$1. Your total earnings will be \$7.

If you choose not to transfer nor take, we will keep the payments unchanged. Your earnings will be \$6.

If you decide to transfer, your payment will be \$3 plus the amount you decide to keep for yourself.

Which action would you like to take?

[Combo box: Take \$1, Neither transfer nor take, Transfer \$1, Transfer \$2, Transfer \$3].

Chapter 3

Morality as Determinant of Social Structure

LUIS AVALOS-TRUJILLO

3.1 Introduction

One of the main longstanding questions in the field of economic development is to explain why some countries are rich while some others are not. Institutions have emerged as a keystone of analysis, not without controversy (Clark, 1987; Diamond, 1997). “Institutions are understood as the humanly devised constraints that shape human interaction” (North, 1990). Despite the indisputable fact that institutions affect economic performance, there has been little analysis devoted to studying how institutions emerge and what are the individual or societal characteristics that shape institutional arrangements. The present study is aimed towards bridging this gap, this article presents a proposal for an experimental procedure to use morality as the founding block of institutions.

The present understanding of the emergence of institutions is wanting. One of the main proposals of the origin of institutions is to place political power and political institutions in a position of predominance (Acemoglu et al. 2005). Economic and political institutions are said to derive from political power (both *de facto* and *de jure*), while political power itself is a result of political institutions. Although elucidating the important role of political power, the argument is mostly circular and does not pin down the origin of institutions in other fundamental unit of analysis beyond the institutions themselves.

One of the most compelling attempts to establish a fundamental basis for the origin of institutions is to be found outside of economics. Boehm (2009), a cultural anthropologist and primatologist, has proposed an explanation for the pervasiveness of egalitarianism in the human species, while it is mostly lacking in other primates. His explanation relies on what he calls a *hierarchy reversal*. Starting from the basis of a hierarchical arrangement as is common in primates, the human species acquired enough dominion over nature leading to an enlargement of the size of the group beyond the sizes common among primates. As the group becomes large, the number of *beta* individuals greatly outnumbers the *alpha*, leading to an egalitarian pull to the access to females, ultimately setting the way for the coevolution of a social norm of egalitarianism and a biological adaptation.

Economics has not been completely absent in this debate. Although not directly stated, it can be inferred from the work of Clark (2008), that norms against violence and institutions for the development of greater literacy and numeracy might also be an example of a coevolution of a social norm and a biological adaptation. In his work, Clark presents substantial evidence that through the centuries prior to the industrial revolution, individuals showed a steady decline for the taste of violence, greater literacy and numeracy, a decline in interest rates which can be attributed to a changing time preference, and an increase of the number of hours worked. Backed by his additional findings that the upper classes (but not the nobility) had substantially more offspring than the lower classes, poses the interesting conclusion that the aforementioned social phenomena were indeed the result of a genetic and normative coevolution. In addition, there has been some theoretical research on the possibility of a coevolution of norms and institutions as shown in the work of Sugden (1998).

The natural step forward is to study the evolution of norms; however, social norms are hardly a fundamental unit of analysis. Norms are usually considered to be the informal rules that govern behavior and they are also thought to be the unplanned result of the individuals' interaction (Bicchieri et al. 2018). For that reason, one ought to consider norms to be a derivative of behavior. More importantly, the existence of a norm presupposes the existence of a notion of "right" (according to the norm) and "wrong" (against the norm). Hence, if one desires to pursue the study of the evolution of norms, the natural course of action is to study the evolution of morality.

Which leads to the main conjecture supporting the present research: institutions

(political and economic) are shaped by our notions of right and wrong, therefore placing morality as the fundamental unit in the analysis of institution formation. At first glance it might seem inappropriate to use morality in economic analysis, however recent research has shown the importance of inequity (Fehr and Schmidt (1999); Bolton and Ockenfels (2000)), fairness (Rabin, 1993; Fehr and Gächter, 2000)¹, honesty (Gneezy et al., 2013, 2018; Hurkens and Kartik, 2009; Lundquist et al., 2009), gratitude and resentment (Smith and Wilson (2019b, 2017)) all of which are intrinsically moral concepts or primitive moral sentiments. Therefore it is fitting to expand and formalize the application of morality to economic analysis.

The aim of this research is to prove the conjecture by means of an experiment. The experimental design rests on using morality as the independent variable while the dependent variable is a social rule or informal constraint that the experimental subjects place upon themselves. If we intend to use morality as the independent variable, one choice is to induce or highlight a set of moral values in the laboratory. Such approach is limiting since the efficacy and persistence of such interventions is not without problems (Russo et al., 2022). Alternatively, we can use pre-existing differences in moral valuations, separate subjects into distinct pools and use an experiment to obtain differentiated outcomes from each pool of subjects. The proposal rests on using the pre-existing heterogeneity in a sample of subjects from a university in the United States, by means of the Moral Foundations questionnaire (Haidt and Joseph, 2004, 2007), which has effectively uncovered the distinct moral motivations of “conservative” and “liberal” subjects.

In the following, we will use the term *binding* to denote those subjects who place an even weight on the five moral foundations: (i) care/harm, (ii) fairness/cheating, (iii) loyalty/betrayal, (iv) authority/subversion and (v) sanctity/degradation. It has been shown that subjects that identify themselves as “conservatives”, follow this pattern. On the other hand, we will use the term *individualizing* to refer to the subjects who place a higher emphasis on the first two while disregarding the rest. We will use the term binding and individualizing since we want to emphasize the distinct moral attitudes and not their political preference.

Once the subjects have been divided into two pools according to their moral system, each pool will engage in what we call the surplus game. The surplus game is meant to be played by a group of 5 experimental subjects who have to decide who among themselves will distribute the surplus that they generated in a previous phase.

¹For comprehensive reviews see (Cappelen and Tungodden, 2019; Fehr and Schmidt, 2006)

There are certain rules: each subject generates surplus by engaging in an effort task while also generating a payoff for themselves, the surplus is differentiated so that each subject cannot “consume” the surplus that she generated. Furthermore, every subject generates a different amount of surplus. The game must be played several rounds and the objective is to determine a rule for the distribution of the surplus across different rounds. It is important to mention that this objective will not be stated explicitly, however it will be stated implicitly since a failure to reach agreement about who will execute the distribution will result in the loss of the surplus. Experimental evidence suggests that binding subjects act less pro-socially in public goods games (Clark et al., 2017; Grünhage and Reuter, 2020), the present design fixes the individual contributions to the surplus.

Since the game is designed to generate different payoffs for each player and a different contribution to the surplus, the game poses an inherent inequity that will likely be overcome by subjects across the moral divide, specially given the ample evidence on inequity aversion (Bellemare et al., 2008; Rey-Biel, 2008; Yang et al., 2016). However, the manner in which this inequity will be overcome is the main purpose of the study. The experiment rests on the evidence that individualizing subjects place a lower relevance to authority (and possibly also hierarchy); furthermore, individualizing subjects place a higher relevance to fairness, which is likely to transpire in a higher proportion of egalitarian rules to elect a steward to distribute the surplus in individualizing groups in contrast to binding groups. Since the main mechanism behind the drive to distribute the surplus is inequity aversion it is important to include additional controls to identify the degree of aversion in each group; we will use the simple task designed by Koch et al. (2021) and use the average of the individual measurements as group measurement.

Since values and morality are usually taken to be somewhat similar, and since there is an overall lack of a conceptual framework to operationalize both concepts in economic research, it is notable that the economics profession is not deprived of an operationalizable definition. Ben-Ner and Putterman (1998) cleverly defined values as *process-regarding preferences*, depriving them of moral content and providing an adequate functional definition. Unfortunately this definition is too broad for the purposes of this research since it encompasses not only our ordinary notions of values as the pillars of morality, but also englobes *proto-institutional* arrangements. For example, a preference for buying an object rather than stealing it is a process-regarding preference; however, electing leaders through periodical voting systems, instead of enduring autocracy and an eventual deposition, is also a process-regarding prefer-

ence. While the former can be conceived under the umbrella of morality, the latter is more akin to a discussion on institutions. The present research can be understood as an attempt to establish a correlation between different moral systems and different proto-institutional arrangements, both of which are process-regarding preferences. We will also refer to these proto-institutional arrangements as *social structure* indistinctly.

The structure of the article is as follows: section 3.2 presents a discussion on the literature related to the present study, the moral foundations theory is presented there in greater length as well as some of its recent applications in economics. Also, that section presents some of the economic literature related with the present research. Section 3.3 contains the proposal for the experimental design as well as the controls to be implemented for further analysis. Section 3.4 briefly explains the proposed estimation. Lastly, a brief discussion on experimental variations and further research is reserved to section 3.5

3.2 Discussion on Related Literature

The proposed experiment is part of the growing literature on the applications of Moral Foundations Theory. For completeness, we will make a succinct but comprehensive overview of the theory. Moral foundations theory was introduced in a series of papers by the pioneering work of Jonathan Haidt and coauthors (Haidt and Joseph, 2004, 2007; Graham et al., 2009). The theory claims that human notions of morality evolved from five psychological modules: (i) care/harm, (ii) fairness/cheating, (iii) loyalty/betrayal, (iv) authority/subversion and (v) sanctity/degradation. Furthermore, the theory claims that each of these modules evolved as an adaptation to the social and natural environment of early humans. It is important to mention that the Theory of Moral Foundations is not the first effort in the study of the origins of morality, since it can be understood as an improvement and refinement of the three ethic systems of Shweder et al. (2013). In that tradition of research, a series of scenarios highlighting different moral dilemmas are used to provide an individual score that qualifies the adherence of the subject to each of the “systems”. In the case of Haidt’s theory, each subject will have a score describing his adherence to each of the five foundations.

In (Haidt and Graham, 2007), the authors apply the theory to describe the moral profile of “conservative” and “liberal” (or “progressive”) subjects in a sample from

the United States. Results show that liberal subjects score high on the first two foundations: care and fairness. Meanwhile, conservative subjects have a medium to high score across all the five foundations. This observation becomes central to our research, since we will explore the economic implications of the two moral systems. Other terms are used to describe each of the moral systems according to their main characteristics. The system that emphasizes care and fairness is also denoted as “individualizing”, while for subjects who rely equally on the five foundations the term “binding” is also used. The intuition behind the terms is that “binding” subjects place some emphasis on the moral foundations that are oriented to the group like loyalty and authority.

Further studies elaborated on the behavioral differences observed across the two groups. An early application studied the behavior from individuals from both moral systems in the traditional social dilemma games, namely the trust game and prisoner’s dilemma (Clark et al. 2017). Results show that or individualizing subjects show more frequent cooperation in the prisoner’s dilemma, a higher transfer in the trust game (trust) and a higher back transfer (trustworthiness). Their results control for age, race and big-5 personality scores. Enke et al. (2020) identifies an additional “trait” underlying the scores on the five foundations: universalism. Universalism is defined as the extent to which people exhibit the same level of altruism and trust towards strangers as towards in-group members. The article shows correlation of universalism with left and right political positions, elicited through self-identification, and policy preferences, elicited through surveys and a governmental spending exercise. It is also interesting to mention is that Enke et al. (2022) shows that subjects who score high in universalism exhibit more “social distancing”, meaning that they have fewer friends and spend less time with them, shedding some light into the social implications of “individualizing” subjects and their socialization. Also recently Schneeberger and Krupka (2021) explored norm compliance of subjects who are assigned to groups with varying degrees of progressivism, which was elicited through the moral foundations questionnaire. Results verify that if the subject feels more identified with the group, they are more likely to comply with the rule. The study focuses on individualizing subjects in individualizing or binding groups.

The study is relevant to the literature of endogenous institutions, especially those related to endogenizing different aspects of the public goods game. For example, Sutter et al. (2010) finds that allowing the subjects to choose if they want to participate in a public goods game with punishment increases the overall level of cooperation, when contrasting against the regular setup where subjects are forced to engage in

the game. In a related study, [Dal Bó et al. \(2010\)](#) shows that subjects are more cooperative in a sequence of prisoner's dilemma games when they are allowed to choose a policy to punish unilateral defections, the policy is voted in groups of four subjects by simple majority and is enforceable to the interactions of players from the group. To a lesser degree, this research is related to the literature of leadership in public goods games ([Gächter and Renner, 2018](#); [Güth et al., 2004](#); [McCannon, 2018](#); [Wang et al., 2017](#)). However, there is a substantial difference between how leadership has been analyzed in public good games. While the literature has focused on the leader as a "first-mover", i.e. the player who chooses an action ahead of the group, while the present research studies leadership from the perspective of power of redistribution or an entitlement to the public good.

The proposed experiment also adds to the scant literature on endogenous institution formation through experimental methods. The literature showcases two powerful examples of social structures emerging spontaneously from the environment. Through a simplified video-game interface, [Wilson et al. \(2012\)](#) proved that ownership rules can emerge from the environment. The subjects in their study successfully replicate different property rules commonly found in whaling communities in the 18th and 19th centuries, which are dependent on the characteristics of the prey. In a more recent experiment, [Camera et al. \(2020\)](#) presented an experimental design where monetary trade emerges in the presence of coordination obstacles, against a backdrop of non-monetary exchange. Somewhat related to the present study is the role-playing literature in comparative politics, the literature is focused in creating complex scenarios that are played by undergraduate students while adhering to a specific role. The literature is an interesting attempt to abstract complex social processes like the formation of parliamentary coalitions ([Biziouras, 2013](#); [Shellman, 2001](#)) or the transition from dictatorship to democracy ([Jiménez, 2015](#)).

The specific design of the game connects this research to the literature on meritocracy and egalitarianism, two recent studies are [Andre, 2021](#) and [Cappelen et al., 2022](#). The first one shows that most subjects judge merit omitting the circumstances under which merit is attained. The study is carried out experimentally using a real effort task and the experimental subject plays the role of an observer. Also using an observer, the second study shows that uncertainty regarding the true performance of a subject in a task makes the observer more egalitarian.

3.3 Experimental Design

The experiment consists of applying the surplus game to differentiated subject pools, for a predetermined number of rounds which would be known in advance by the participants. The surplus game consists of a social dilemma game where each subject generates surplus units which cannot be consumed or allocated to the subject who produced them. All surplus units must be distributed by one of the participants in every round, and the participant that executes the distribution must be chosen through simple majority. The design highlights the conflict between existing moral tendencies; on one hand the tendency to provide a fair payoff to all participants, and the necessity of a centralized figure who would be able to execute such redistribution. Thus, the game highlights a moral tradeoff between the care and fairness foundation on one side, and the authority foundation on the other. A key element in the game is that each subject is assigned to a task in each round and that each task generates a different surplus and a different payoff to the subject. The objective of the game is to capture the different rules that subjects use to assign leadership and distribute the surplus.

The surplus game consists of three stages: production phase, deliberation phase and distribution phase which we now explore in detail.

3.3.1 Production Phase

In the production phase each subject is assigned randomly to a task. Tasks are numbered 1 through 5 and they correspond to different levels of effort required to complete the task. The task consists in clicking all the zeros in a sequence of squared matrices containing zeros and ones. Participants are not allowed to advance to the next matrix unless they have clicked all the zeros, and only the zeros in the matrix at hand. For a schematic representation refer to figure [3.3.1](#). The size and number of the matrices varies according to the index of the task. Once the subject finishes her task she earns a given number of tokens for herself and a given number of tokens to the surplus. The number of tokens that are awarded to the player and to the surplus depend on the index of the task.

The number of tokens generated for the player and as surplus is summarized in table [3.3.1](#). The table also indicates the appropriate level of effort needed to complete the task, which is done without time limit. In order to provide an easy way

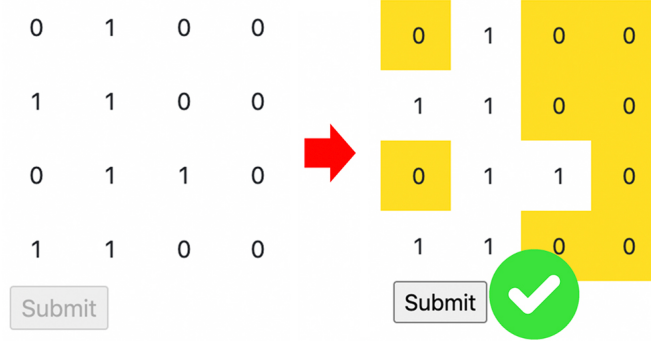


Figure 3.3.1: Schematic representation of the real effort task.

Task ID	Color Code	Task (#, dimension)	Tokens for the Player	Surplus Tokens
1	Red	5, 15×15	5	1
2	Blue	7, 12×12	4	2
3	Green	7, 10×10	3	3
4	Purple	7, 8×8	2	4
5	Yellow	10, 5×5	1	5

Table 3.3.1: Payoffs and surplus summary by task. All tasks consist in clicking the zeros on a matrix populated by zeros and ones.

for the subjects to distinguish each task, and to provide a sense of entitlement (or ownership) to the tokens generated (Wilson et al., 2012), each task is color coded. Color code information can also be consulted in table 3.3.1. For example, task 2 consists of 7 matrices of 12×12 and generates 6 blue tokens, 2 tokens are for the surplus and 4 tokens are for the player.

Note that every subject generates the same number of tokens but the experimental design induces some inequality among the players. The goal of the game is to overcome the inherent inequality by means of the distribution phase. Since each subject cannot be allocated the surplus units she generated participants must agree on a convenient process to reassign the tokens.

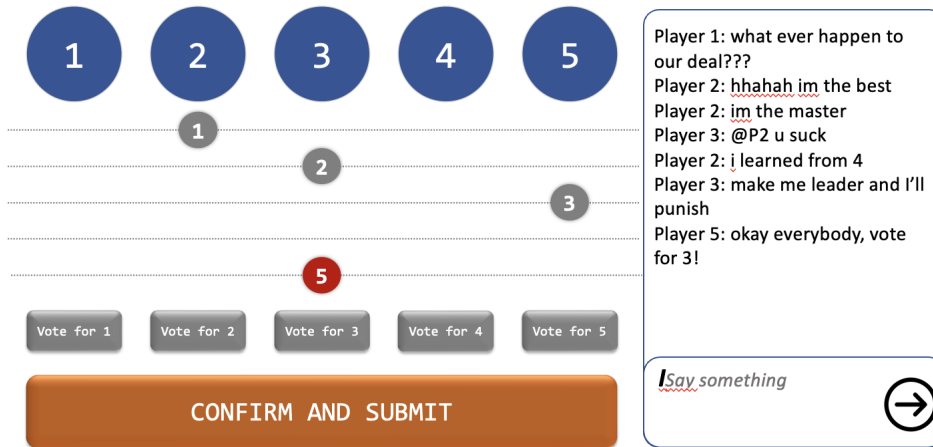


Figure 3.3.2: Representation of the graphic interface used for the deliberation phase.

3.3.2 Deliberation Phase

Once all subjects have completed their allotted task they proceed to the deliberation phase. The deliberation phase is carried out through a graphic interface containing two elements: a voting system and a chat. The voting system is designed to allow each subject to vote for which participant will carry out the distribution (it is possible for a subject to vote for herself), but also to inform all other subject the voting intentions of all other participants. Voting is carried out in two steps, first the subject reveals a voting intention which is made public to all other players; in this stage the subject can revise her vote as many times as needed. Once the subject has reached a decision on who to vote for, the subject can lock-in her vote. The vote becomes permanent and all other subjects are informed. Adjacent to the voting system, subjects have a free-form chat which they can use to discuss who will execute the distribution and also how to execute it. Note that since this is a free-form chat all forms of self-expression are allowed. It is possible that subjects could also use this tool to promote themselves or to ask for punishments to be carried out. The hypothesis is that subjects will use this free-form chat to come up with a rule to elect the leaders and also a redistribution rule. If a rule emerges within the group it will be coded as success, and the type of rule will also be coded. We will elaborate more on coding the rule in section [3.4.1](#). The graphic interface for the deliberation phase is shown in figure [3.3.2](#).

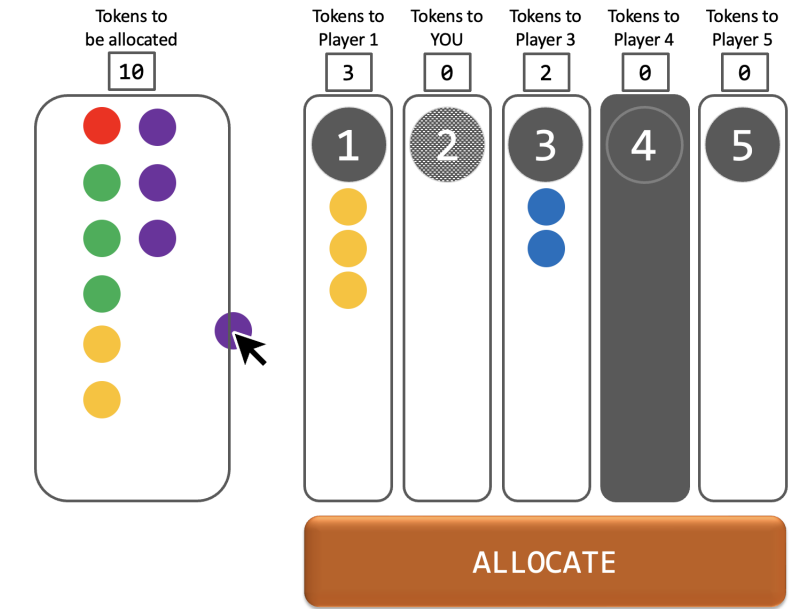


Figure 3.3.3: Representation of the graphic interface used for the distribution phase. Note that purple tokens cannot be assigned to subject 4, hence that option is blocked out.

3.3.3 Distribution Phase

In the distribution phase, the subject elected as dictator has to distribute the surplus tokens among the group. The interface will be designed to enforce the rule that the tokens cannot be allocated to the subject who produced them. The allocation phase is done in a private manner and all other subjects are informed of the final distribution at the end of this phase. A representation of the graphic interface used for this stage is shown in figure [3.3.3](#). Note that the subject elected as dictator has absolute freedom in how to execute the distribution of the surplus. This raises potential issues of defection or deviation from pre-agreed distributions.

3.3.4 Multiround Surplus Game

The surplus game is intended to be used in multiple rounds. The reason for this is that a one-round implementation will not be enough to uncover the solution of the social dilemma. Indeed the social dilemma induced by this game only emerges when the game is played for multiple rounds. In that case the game hinges on which

subject gets assigned which task in each round, which in turn determines the structure of the inequality induced by the game. For the present design, we will induce inequality through a stochastic process so that each subject is more likely to be assigned the same task in the next round. At inception, each subject is equally likely to be assigned any task, subsequent assignment of tasks will be done favoring the current allocation and placing a lower probability on drastically distinct allocations.

To be more precise, let the tasks be given by the following set $T = \{1, 2, \dots, n\}$ where n is the number of tasks. Note that in our case, the number of players equals the number of tasks, hence the assignment of tasks to subjects can be thought of as a permutation of the set T . If we denote as T_n as all the possible permutations of T in groups of n , then given $r \in T_n$ we have that $r(i)$ is the task assigned to subject i . Let r_0 denote the initial assignment of tasks to each subjects. The allocation algorithm is such that $P(r_0) = P(s_0)$ for any $r_0, s_0 \in T_n$ and $r_0 \neq s_0$. In other words, all permutations are equal likely at inception. In subsequent stages, let $P(r, r_0)$ denote the transition probability from permutation r_0 to permutation r in the next stage, the transition rule is that $P(r, r_0) = xM^d$ for $r \in T_n$ such that $K(r, r_0) = d$ where K denotes the Kendall tau rank distance. To complete the transition rule we need, $P(r, r_0) = P(s, r_0)$ for $r, s \in T_n$ such that $K(r, r_0) = K(s, r_0)$.

Recall that the Kendall tau distance counts the number of disagreements between two ranking lists, equivalently it counts the number of swaps that the bubble sort algorithm would take to place one list in the same order as the other list. In our case, any permutation $r \in T_n$ can be described as an ordered list. The rule governing this stochastic process is that $P(r_0, r_0) = x$, the probability of staying at the same assignment for the next round is x . In further rounds, $P(r, r_0) = xM = MP(r_0, r_0)$ if $K(r, r_0) = 1$, which means that keeping the current assignment is M times more likely than just swapping the assignments between two players. If $K(r, r_0) = 2$ then we need to swaps tasks twice to get from r_0 to r , passing through the intermediate assignment r' , hence $P(r, r_0) = MP(r', r_0) = M^2P(r_0, r_0)$. This means that the current assignment is M^2 times more likely than swapping tasks twice. The discussion clarifies the general rule for governing transition probabilities: from any given assignment of tasks, it is more likely to stay in the current assignment, and a transition to another assignment is more unlikely according to a factor of M^d , where d is the number of swaps. Finally, transitioning to any two assignments that require the same number of swaps is equally likely.

3.3.5 Additional Controls and Questionnaires

As mentioned earlier, the subject pool will be divided into “individualizing” and “binding”. In order to effect that division two instruments will be used: the moral foundations questionnaire and the global universalism question from [Enke et al. \(2020\)](#). The study will employ the July 2008 version of the moral foundations questionnaire which comprises 32 questions in two parts. The first part involves hypothetical qualifiers to a scenario where right or wrong is judged. The subject has to rate how important is the qualifier for her to decide if an action is deemed right or wrong. For example, item 9 states “Whether or not someone conformed to the traditions of society” and the subject must declare if when judging an action to be right or wrong, conforming to traditions of society is relevant or not in a 6 point Likert scale where 1 corresponds to “not at all relevant” and 6 corresponds to “extremely relevant”. The second part of the questionnaire consists of 16 statements where the subject has to declare agreement in 6 point Likert scale where 1 corresponds to “Strongly disagree” and 6 corresponds to “Strongly agree”. The questionnaire includes items such as “I am proud of my country’s history” or “It can never be right to kill a human being”. The full questionnaire is reproduced in appendix [3.A](#). The final step for analysis consists in summarizing the information obtained from the moral foundations questionnaire into an index that would reflect the subject’s adherence to a binding moral matrix. The dimensional reduction can be carried out through principal component analysis or using a variety of clustering techniques.

To reinforce the results obtained from the moral foundations questionnaire we will also use an excerpt from the global universalism questionnaire originally proposed by [Enke et al. \(2020\)](#). According to their study, adherence to a specific moral matrix can be summarized by their adherence to universalism. A subject is said to be a “universalist” when she exhibits the same level of altruism towards subjects regardless of their social distance, this means that a universalist subject will be as altruistic towards a family member than towards a stranger from a foreign country. We will use the questionnaire for “Domestic Universalism in Altruism” and the questionnaire for “Foreign Universalism in Altruism”. Those questionnaires consists of a hypothetical scenario where \$100 are split by the subject between a random person from the United States and a different subject that could be either a relative or a member of the same church or a random person from another country. The questionnaires are reproduced in appendix [3.B](#) and they will be applied as a single questionnaire for the purposes of this research. In addition, a measure of universalism will be created by subtracting the responses between the two most extreme measures: the division between a random person from the U.S. and a friend of a family member (denoted

f), against the division between a random person from the U.S. a random person who lives anywhere in the world (denoted w). Hence the measure of universalism u is defined as $f - w$. A perfect universalist will exhibit a value of u equal to zero; positive values of u denote greater degree of deviation from universalism. Alternatively, we can use the composite measure of universalism in altruist by using the unweighted average of the responses in each of the items, which was the composite measure employed by the original authors of this questionnaire.

Since the experiment relies heavily on personal notions of fairness, it is important to provide an additional control for inequity aversion (Fehr and Schmidt (1999); Bolton and Ockenfels (2000)). Inequity aversion was originally measured via dictator and ultimatum games, however the interpretation of those games as indicators of inequity aversion has been highly disputed (Engelmann and Strobel 2004; Fershtman et al. 2012; List 2007). Since our study would require a quick measurement, using an ultimatum game prior to the main experiment might hinder the main objective of this study. For that reason this study opt to use the proposed measurement by Koch et al. (2021) that relies on a simple incentivized question in the form of a coupled lottery. In this method, two subjects engage in identical binary lotteries (they earn M with probability p or zero with probability $1 - p$), their choice consists in deciding if they want to take on this lottery independently or coupled with another participant. If they choose to take the lottery independently, the outcome of the lotteries will be decided by two independent random draws and they payouts will be given to each player. If the player decided to play coupled, then the outcome will be decided by just one random draw that will decide that payouts for both players. It is important to mention that each subject faces the same decision, so there is a chance that the subjects do not agree on how they would like to play. In the case of conflicting responses, a random draw will decide if the payouts are carried out independently or coupled. For the present study, the coupled lottery will be implemented prior to the surplus game. All subjects will be informed that some participants will receive an additional payoff from the coupled lottery, those subjects will be chosen randomly.

3.3.6 Summary of Experimental Procedure

The experimental procedure can be summarized as follows

- i. Apply the moral foundations questionnaire and the universalism questionnaires.

- ii. Separate the subject pool in two: binding and individualizing.

At a later date, in each experimental session:

- i. Subjects answer the inequity aversion questionnaire individually
- ii. Subjects are grouped in teams of 5 players and they play 5 rounds of the surplus game
- iii. Subjects answer the demographic questionnaire individually
- iv. Subjects are paid according to their choices

3.4 Analysis and Estimation

3.4.1 Outcomes, Rules and Hypotheses

Given the complexity of the game, a clearcut prediction is impossible to attain. However, previous literature in prosocial behavior inform us of possible outcomes. As mentioned in the introduction, the dependent variable to be analyzed is not the outcome but the process through which the outcome is achieved. Hence, hypotheses on the possible rules to solve the social dilemma are also required. Outcomes in each round will be classified under three labels: failure, success and defection. Failure means that subjects failed to agree or did not want to agree on a leader which results in the loss of the surplus. Success means that subjects agreed on a leader and some distribution of the surplus was made. Defection means that the subject elected as leader keep all or most the surplus to herself, crucial to defection will be the presence of a punishment which is expected in the form of ostracism (Liddell and Kruschke, 2014; Maier-Rigaud et al., 2010). In addition, we will track how the surplus is distributed; the number of competitive elections, which means elections that are decided 3 vs 2; and number of surplus tokens that is wasted across all rounds.

Processes or rules will be classified in two groups according to their relationship to the stochastic process generating the structural inequality. The rule will be said to be *pro-stochastic* if the choice of leaders follows the stochasticity in the process. For example, if the rule is that the leader should be the subject who contributed the most to the surplus, then the rule is pro-stochastic since it follows the underlying stochastic process assigning the tasks. On the contrary, if the rule that emerges

makes no reference to the stochastic process assigning the tasks it is said to be *anti-stochastic*. Some examples of anti-stochastic rules are: to choose whomever had the best distribution proposal or to choose the leader before the uncertainty is resolved.

The hypothesized outcomes and rules depend heavily on the behavioral assumptions on the subjects. There are three main sets of assumptions which will be elaborated below. In all cases we assume common knowledge.

i) **Selfish:** All subjects are selfish.

Hence the leader will also be selfish and will steal all the surplus. Under common knowledge, a selfish subject knows that the subject assigned with task 5 (yellow tokens) is the one who can steal less since she produced the most tokens for the surplus; it is rational to elect that subject. The leader will steal all but 5 tokens and distribute one yellow token to each subject. In this case the rule is pro-stochastic.

ii) **Inequality Averse:** All subjects are inequality averse.

The leader will also be inequality averse. Under reasonable parametric assumptions using the most common inequality aversion models (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), the leader will distribute the surplus so that the total earnings in each round equals 6 tokens. Under common knowledge, the outcome is not an issue but the rule must still be established. There are two possibilities:

A. Under fairness or egalitarian concerns the rule will be to choose the leader in advance of the task assignment. This rule corresponds to a choice behind the veil of ignorance (Dworkin, 1981; Harsanyi, 1953; Rawls, 2020) and it will be classified as anti-stochastic.

B. Under meritocratic concerns the leader is chosen in relationship with her contribution to the surplus (task 5, yellow tokens) or for individual merit (task 1, red tokens). The rule is then classified as pro-stochastic.

iii) **Social Selfish:** The leader is selfish but the rest of subjects are inequality averse.

In this case we can expect a well known behavior in ultimatum games and public good games with punishment. The subject elected as leader will prefer to redistribute rather than steal the surplus and face punishment, most likely via ostracism. This means that a leader who defects will most likely be omitted in the distribution for the next period or perhaps even for the entire game. In

this case, the leader will distribute the surplus. Note that if we assume that all subjects are inequality averse but once elected leader the subject becomes selfish, then the choice of leader is completely irrelevant as the surplus will be distributed nonetheless. Therefore we will observe an anti-stochastic rule in this case. Note that despite the rule being anti-stochastic, this rule involves no fairness considerations since the outcome before or after the veil of ignorance is identical.

The objective of this study is to establish a relationship between the rules and the underlying moral matrix of the group. The working hypothesis in this study is that binding groups will be more amenable to rules that preserve some hierarchical inequality among the subjects. In the same direction, individualizing groups will place more emphasis on leveling off the inequality induced by the stochastic assignment of tasks.

Hypothesis 1. Individualizing groups will be more likely to select anti-stochastic rules and binding groups will be more likely to select pro-stochastic rules.

However, if the meritocratic concerns of the binding groups prevails over the inequality aversion concerns it is reasonable to expect them to block any individual attempt to distribute the surplus. There are sufficient reasons to believe that binding groups will fail to distribute the surplus since conservative individuals are less likely to be in favor of distributive policies as evidenced in surveys [Enke et al. \(2020\)](#) and experimental interventions [\(Grünhage and Reuter, 2020\)](#) albeit with some limitations [\(Anderson et al., 2005\)](#). In addition, individualizing subjects have been shown to be more prosocial in certain social dilemma games like the trust game or the prisoner's dilemma [\(Clark et al., 2017\)](#). [conservatives will have a higher sense of entitlement to their tokens].

Hypothesis 2. Binding groups will fail to distribute the surplus more often than individualizing groups.

3.4.2 Proposed Estimation Procedures

The variable of interest is if the rule that each group decides for the groups is anti-stochastic or not, it will be denoted by the binary variable A . We want to uncover the relationship between the rule and the social characteristics of the group, mainly if the group is composed of binding or individualizing subjects. We will use the binary

variable *bind* to denote if the group is binding. In addition, we want to control for the effect of inequity aversion of the group. Inequity aversion is measured at the individual level by means of their response to the coupled lottery question. Recall that each subject will face a binary lottery where they can earn M with probability $p = 0.5$ or nothing with probability $1 - p = 0.5$. The subject is paired randomly with another subject in the same session and he faces the decision of coupling the lottery with his partner or playing independently. If both subjects choose to play independently, two random draws decide the outcome for each subject. If both subjects choose to couple their lotteries, then one random draw will decide the payoffs for both players. If they disagree then the outcome will be randomly decided between coupled or independent. Note that an inequity averse subject would choose to couple the lotteries and will be coded as $q_i = 1$. The inequity aversion of the group will be measured as the simple average of the binary choices ($\text{ineq} = 1/n \sum_i q_i$), thus denoting the proportion of subjects who coupled their lotteries. The estimation is carried out by means of the following equation:

$$P(A = 1) = \Lambda(\beta_0 + \beta_1 \text{bind} + \beta_2 \text{ineq} + u) \quad (3.1)$$

where Λ denotes the logit function. An alternative specification is to substitute the term *bind* by the universalism measure *univ*, which is not binary.

3.5 Concluding Remarks

In the preceding sections we have described an experimental procedure to describe how morality (or value systems) transpires into greater societal constructs, which we have called *proto-institutions* or social constructs. The experiment rests on separating the subject pool in two groups according to their adherence to either binding or individualizing moral matrices. The experiment itself consists of grouping the subjects into teams of 5, each subject must tackle a series of effort tasks to produce a payoff for themselves and a fixed contribution to the surplus. Individual payoffs and contributions to the surplus are heterogeneous. Their main task is to agree to select one participant among them who will distribute the surplus at will. The distribution is preceded by a deliberation phase where subjects can freely chat with each other and cast votes publicly.

The present proposal could be further simplified by reducing the size of the group to 3 and by simplifying the induced inequality. Instead of giving a differentiated pay-

off and contribution to the surplus, 2 subjects could earn the same individual payoff and provide the same contribution to the surplus while one subject has a greater personal payoff than the other two and a smaller contribution to the surplus. Another interesting variation is to exacerbate the inequality in the experimental design by giving a higher individual payoff to subjects who have a higher contribution to the surplus, the current design runs exactly the opposite. Such intervention is of particular interest because it will induce a sense of hierarchy, which is one of the points of contention between binding and individualizing subjects. Also, this might be a cleaner way to induce anti-stochastic choices among the individualizing population and perhaps a higher proportion of pro-stochastic choices among the binding subjects. The main danger of such design is that egalitarian concerns might be the primary motivation for action in both subpopulations, rendering ineffectual any attempt to measure a differentiated behavior.

Finally, regardless of the specific experimental design, the study could include an additional treatment where subjects from both subpopulations are present in the same group. This would be an interesting contrast to both subpopulations and an interesting exercise of external validity.

Appendix

3.A Moral Foundations Questionnaire

Part 1.

When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

0 - not at all relevant (This consideration has nothing to do with my judgments of right and wrong)

1 - not very relevant

2 - slightly relevant

3 - somewhat relevant

4 - very relevant

5 - extremely relevant (This is one of the most important factors when I judge right and wrong)

- Whether or not someone suffered emotionally
- Whether or not some people were treated differently than others
- Whether or not someone's action showed love for his or her country
- Whether or not someone showed a lack of respect for authority
- Whether or not someone violated standards of purity and decency
- Whether or not someone was good at math
- Whether or not someone cared for someone weak or vulnerable
- Whether or not someone acted unfairly

- Whether or not someone did something to betray his or her group
- Whether or not someone conformed to the traditions of society
- Whether or not someone did something disgusting
- Whether or not someone was cruel
- Whether or not someone was denied his or her rights
- Whether or not someone showed a lack of loyalty
- Whether or not an action caused chaos or disorder
- Whether or not someone acted in a way that God would approve of

Part 2.

Please read the following sentences and indicate your agreement or disagreement:

0 - Strongly disagree

1 - Moderately disagree

2 - Slightly disagree

3 - Slightly agree

4 - Moderately agree

5 - Strongly agree

- Compassion for those who are suffering is the most crucial virtue.
- When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
- I am proud of my country's history.
- Respect for authority is something all children need to learn.
- People should not do things that are disgusting, even if no one is harmed.
- It is better to do good than to do bad.
- One of the worst things a person could do is hurt a defenseless animal.
- Justice is the most important requirement for a society.
- People should be loyal to their family members, even when they have done something wrong.

- Men and women each have different roles to play in society.
- I would call some acts wrong on the grounds that they are unnatural.
- It can never be right to kill a human being.
- I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.
- It is more important to be a team player than to express oneself.
- If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.
- Chastity is an important and valuable virtue.

3.B Universalism Questionnaire

In each row below, how would you split 100 between a randomly selected person who lives in the United States and the individual displayed on the right (who is part of a particular social group)?

The closer you drag the slider to one individual, the more money you allocate to that individual. Please assume all individuals below have the same income, all live in the United States, and I will not find out that it was you who sent them the money.

The interface shows 6 sliders, on the right it always says: "Randomly-selected person [index] who lives in the United States". On the left it lists the following:

- A friend of a family member (e.g. your sibling's closest friend)
- A member of your extended family (e.g. your cousin)
- Former or current colleague at work or school
- Someone who shares your religious beliefs (e.g. a fellow Christian)
- A member of one of your pastor current organizations (local church, leisure club, etc.)
- A randomly selected person who lives anywhere in the world

Bibliography

- Acemoglu, D., Johnson, S., and Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of Economic Growth*, 1:385–472.
- Algan, Y. and Cahuc, P. (2010). Inherited trust and growth. *American Economic Review*, 100(5):2060–2092.
- Anderson, L. R., Mellor, J. M., and Milyo, J. (2004). Social capital and contributions in a public-goods experiment. *American Economic Review*, 94(2):373–376.
- Anderson, L. R., Mellor, J. M., and Milyo, J. (2005). Do liberals play nice? the effects of party and political ideology in public goods and trust games. In *Experimental and Behavioral Economics*. Emerald Group Publishing Limited.
- Andersson, O., Holm, H. J., Tyran, J.-R., and Wengström, E. (2016). Deciding for others reduces loss aversion. *Management Science*, 62(1):29–36.
- Andre, P. (2021). Shallow meritocracy: An experiment on fairness views. Technical report, ECONtribute Discussion Paper.
- Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388.
- Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.
- Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.
- Bartlett, M. Y. and DeSteno, D. (2006). Gratitude and prosocial behavior: Helping when it costs you. *Psychological Science*, 17(4):319–325.

- Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., and Todd, R. M. (1991). Empathic joy and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, 61(3):413.
- Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., and Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, 55(1):52.
- Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.
- Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.
- Beeler-Duden, S. and Vaish, A. (2020). Paying it forward: The development and underlying mechanisms of upstream reciprocity. *Journal of Experimental Child Psychology*, 192:104785.
- Bellemare, C., Kröger, S., and Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.
- Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.
- Ben-Ner, A. and Putterman, L. (1998). Values and institutions in economic analysis. In Ben-Ner, A. and Putterman, L., editors, *Economics, values and organization*, chapter 1, pages 3–69. Cambridge University Press, Cambridge, United Kingdom.
- Berkowitz, L. and Daniels, L. R. (1964). Affecting the salience of the social responsibility norm: effects of past help on the response to dependency relationships. *The Journal of Abnormal and Social Psychology*, 68(3):275.
- Bicchieri, C., Muldoon, R., and Sontuoso, A. (2018). Social Norms. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition.
- Biziouras, N. (2013). Midshipmen form a coalition government in belgium: Lessons from a role-playing simulation. *PS: Political Science & Politics*, 46(2):400–405.

- Boehm, C. (2009). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.
- Bolton, G. E., Katok, E., and Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics*, 89(8):1457–1468.
- Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 91(1):166–193.
- Bowman, D., Minehart, D., and Rabin, M. (1999). Loss aversion in a consumption–savings model. *Journal of Economic Behavior & Organization*, 38(2):155–178.
- Boyd, R. and Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.
- Boyd, R. and Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks*, 11(3):213–236.
- Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.
- Breitmoser, Y. and Tan, J. H. (2013). Reference dependent altruism in demand bargaining. *Journal of Economic Behavior & Organization*, 92:127–140.
- Buckenmaier, J. and Dimant, E. (2021). The experience is (not) everything: Sequential outcomes and social decision-making. *Economics Letters*, page 109916.
- Cabral, L., Ozbay, E. Y., and Schotter, A. (2014). Intrinsic and instrumental reciprocity: An experimental study. *Games and Economic Behavior*, 87:100–121.
- Camera, G., Goldberg, D., and Weiss, A. (2020). Endogenous market formation and monetary trade: an experiment. *Journal of the European Economic Association*, 18(3):1553–1588.
- Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441.
- Cappelen, A. W., Mollerstrom, J., Reme, B.-A., and Tungodden, B. (2022). A meritocratic origin of egalitarian behaviour. *The Economic Journal*, 132(646):2101–2117.

- Cappelen, A. W. and Tungodden, B. (2019). *The economics of fairness*. Edward Elgar Publishing Limited.
- Chagnon, N. A. (1988). Life histories, blood revenge, and warfare in a tribal population. *Science*, 239(4843):985–992.
- Chancellor, J., Margolis, S., Jacobs Bao, K., and Lyubomirsky, S. (2018a). Everyday prosociality in the workplace: The reinforcing benefits of giving, getting, and glimpsing. *Emotion*, 18(4):507.
- Chancellor, J., Margolis, S., and Lyubomirsky, S. (2018b). The propagation of everyday prosociality in the workplace. *The Journal of Positive Psychology*, 13(3):271–283.
- Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.
- Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9(C):88–97.
- Clark, C. B., Swails, J. A., Pontinen, H. M., Bowerman, S. E., Kriz, K. A., and Hendricks, P. S. (2017). A behavioral economic assessment of individualizing versus binding moral foundations. *Personality and Individual Differences*, 112:49–54.
- Clark, G. (1987). Why isn't the whole world developed? lessons from the cotton mills. *The Journal of Economic History*, 47(1):141–173.
- Clark, G. (2008). A farewell to alms. In *A Farewell to Alms*. Princeton University Press.
- Cooper, D. J. and Kagel, J. H. (2016). Other-regarding preferences. *The Handbook of Experimental Economics*, 2:217.
- Cosaert, S., Lefebvre, M., and Martin, L. (2022). Are preferences for work reference dependent or time nonseparable? new experimental evidence. *European Economic Review*, 148:104206.

- Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5):2205–2229.
- Dana, J., Cain, D. M., and Dawes, R. M. (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, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.
- Dawes, C. T., Johannesson, M., Lindqvist, E., Loewen, P. J., Ostling, R., Bonde, M., and Priks, F. (2012). Generosity and political preferences.
- DeSteno, D., Bartlett, M. Y., Baumann, J., Williams, L. A., and Dickens, L. (2010). Gratitude as moral sentiment: emotion-guided cooperation in economic exchange. *Emotion*, 10(2):289.
- Diamond, J. (1997). *Guns, germs and steel*. W.W. Norton.
- Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.
- Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.
- Dufwenberg, M., Gneezy, U., Güth, W., and Van Damme, E. (2001). Direct vs indirect reciprocity: an experiment. *Homo Oecon*, 18:19–30.
- Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.
- Dworkin, R. (1981). What is equality? part 2: Equality of resources. *Philosophy & Public Affairs*, pages 283–345.
- Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.
- Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869.

- Enke, B., Rodríguez-Padilla, R., and Zimmermann, F. (2020). Moral universalism and the structure of ideology. Technical report, National Bureau of Economic Research.
- Enke, B., Rodríguez-Padilla, R., and Zimmermann, F. (2022). Moral universalism: Measurement and economic relevance. *Management Science*, 68(5):3590–3603.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.
- Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness—intentions matter. *Games and Economic Behavior*, 62(1):287–303.
- Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.
- Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–182.
- Fehr, E. and Gintis, H. (2007). Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology*, 33:43–64.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.
- Fehr, E. and Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism—experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity*, 1:615–691.
- Fershtman, C., Gneezy, U., and List, J. A. (2012). Equity aversion: Social norms and the desire to be ahead. *American Economic Journal: Microeconomics*, 4(4):131–144.
- Fosgaard, T. R., Hansen, L. G., and Wengström, E. (2019). Cooperation, framing, and political attitudes. *Journal of Economic Behavior & Organization*, 158:416–427.
- Füllbrunn, S. C. and Luhan, W. J. (2017). Decision making for others: The case of loss aversion. *Economics Letters*, 161:154–156.
- Gächter, S. and Renner, E. (2018). Leaders as role models and ‘belief managers’ in social dilemmas. *Journal of Economic Behavior & Organization*, 154:321–334.

- Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.
- Gilens, M. and Thal, A. (2018). Doing well and doing good? how concern for others shapes policy preferences and partisanship among affluent americans. *Public Opinion Quarterly*, 82(2):209–230.
- Glaeser, E. L., Laibson, D., and Sacerdote, B. (2002). An economic approach to social capital. *The Economic Journal*, 112(483):F437–F458.
- Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–453.
- Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, 93:293–300.
- Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029.
- Greiner, B. and Levati, M. V. (2005). Indirect reciprocity in cyclical networks: An experimental study. *Journal of Economic Psychology*, 26(5):711–731.
- Grünhage, T. and Reuter, M. (2020). Political orientation is associated with behavior in public-goods-and trust-games. *Political Behavior*, pages 1–26.
- Guala, F. (2010). Reciprocity: Weak or strong? what punishment experiments do (and do not) demonstrate. *University of Milan Department of Economics, Business and Statistics Working Paper*, (2010-23).
- Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, 55(1):25–30.
- Guiso, L., Sapienza, P., and Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94(3):526–556.
- Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *The Quarterly Journal of Economics*, 124(3):1095–1131.
- Guiso, L., Sapienza, P., and Zingales, L. (2011). Civic capital as the missing link. *Handbook of Social Economics*, 1:417–480.

- Güth, W., Levati, M. V., Sutter, M., and van der Heijden, E. (2004). Leadership and cooperation in public goods experiments. *Max Planck Institute of Economics*.
- Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., J., D.-M., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. (2020). World values survey: Round seven.
- Haidt, J. and Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.
- Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.
- Haidt, J. and Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind*, 3:367–391.
- Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(5):434–435.
- Haun, D. B. and Tomasello, M. (2011). Conformity to peer pressure in preschool children. *Child Development*, 82(6):1759–1767.
- Haushofer, J., Biletzki, A., and Kanwisher, N. (2010). Both sides retaliate in the israeli–palestinian conflict. *Proceedings of the National Academy of Sciences*, 107(42):17927–17932.
- Horowitz, D. (1985). *Ethnic Groups in Conflict*. University of California Press.
- Hugh-Jones, D., Ron, I., and Zultan, R. (2019). Humans reciprocate by discriminating against group peers. *Evolution and Human Behavior*, 40(1):90–95.
- Hurkens, S. and Kartik, N. (2009). Would i lie to you? on social preferences and lying aversion. *Experimental Economics*, 12:180–192.
- Isen, A. M. and Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology*, 21(3):384.
- Jiménez, L. F. (2015). The dictatorship game: Simulating a transition to democracy. *PS: Political Science & Politics*, 48(2):353–357.
- Kant, I. (1788:2002). *Critique of practical reason*. Hackett Publishing.

- Kawagoe, T. and Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102:1–9.
- Kerschbamer, R. and Müller, D. (2020). Social preferences and political attitudes: An online experiment on a large heterogeneous sample. *Journal of Public Economics*, 182:104076.
- Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.
- Kimbrough, E. and Vostroknutov, A. (2019). A theory of injunctive norms. mimeo.
- Koch, M., Menkhoff, L., and Schmidt, U. (2021). Coupled lotteries—a new method to analyze inequality aversion. *Journal of Economic Behavior & Organization*, 191:236–256.
- Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.
- Liddell, T. M. and Kruschke, J. K. (2014). Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. *Judgment and Decision Making*, 9(6):523–547.
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3):482–493.
- Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, 70(1-2):81–92.
- Maier-Rigaud, F. P., Martinsson, P., and Staffiero, G. (2010). Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior & Organization*, 73(3):387–395.
- McCabe, K. A., Rigdon, M. L., and Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, 52(2):267–275.
- McCannon, B. C. (2018). Leadership and motivation for public goods contributions. *Scottish Journal of Political Economy*, 65(1):68–96.
- Mitzkewitz, M. and Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, 22(2):171–198.

- Mujcic, R. and Leibbrandt, A. (2018). Indirect reciprocity and prosocial behaviour: evidence from a natural field experiment. *The Economic Journal*, 128(611):1683–1699.
- Naef, M. and Schupp, J. (2009). Can we trust the trust game? a comprehensive examination. *Royal Holloway College, Discussion Paper Series*, 5.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- Nowak, M. A. and Roch, S. (2007). Upstream reciprocity and the evolution of gratitude. *Proceedings of the royal society B: Biological Sciences*, 274(1610):605–610.
- Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577.
- Opp, K.-D. (1982). The evolutionary emergence of norms. *British Journal of Social Psychology*, 21(2):139–149.
- Over, H. and Carpenter, M. (2012). Putting the social into social learning: explaining both selectivity and fidelity in children’s copying behavior. *Journal of Comparative Psychology*, 126(2):182.
- Polman, E. (2012). Self–other decision making and loss aversion. *Organizational Behavior and Human Decision Processes*, 119(2):141–150.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, pages 1281–1302.
- Rawls, J. (2020). *A theory of justice: Revised edition*. Harvard university press.
- Rey-Biel, P. (2008). Inequity aversion and team incentives. *scandinavian Journal of Economics*, 110(2):297–320.
- Roemer, J. E. (2010). Kantian equilibrium. *Scandinavian Journal of Economics*, 112(1):1–24.
- Romano, A., Saral, A. S., and Wu, J. (2021). Direct and indirect reciprocity among individuals and groups. *Current Opinion in Psychology*.

- Ross, L., Greene, D., and House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.
- Russo, C., Danioni, F., Zagrean, I., and Barni, D. (2022). Changing personal values through value-manipulation tasks: a systematic literature review based on schwartz’s theory of basic human values. *European Journal of Investigation in Health, Psychology and Education*, 12(7):692–715.
- Sapienza, P., Toldra-Simats, A., and Zingales, L. (2013). Understanding trust. *The Economic Journal*, 123(573):1313–1332.
- Schmidt, M. F., Butler, L. P., Heinz, J., and Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, 27(10):1360–1370.
- Schneeberger, A. and Krupka, E. L. (2021). Determinants of norm compliance: Moral similarity and group identification. *Available at SSRN 3969227*.
- Schroeder, D. A., Dovidio, J. F., Sibicky, M. E., Matthews, L. L., and Allen, J. L. (1988). Empathic concern and helping behavior: Egoism or altruism? *Journal of Experimental Social Psychology*, 24(4):333–353.
- Schwerter, F. and Zimmermann, F. (2020). Determinants of trust: the role of personal experiences. *Games and Economic Behavior*, 122:413–425.
- Selten, R. (1967). Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopol-experiments, s. 136–168. *Beiträge zur experimentellen Wirtschafts-Forschung*. Tübingen: JCB Mohr.
- Sen, S. and Airiau, S. (2007). Emergence of norms through social learning. In *IJCAI*, volume 1507, page 1512.
- Shayo, M. and Zussman, A. (2011). Judicial ingroup bias in the shadow of terrorism. *The Quarterly Journal of Economics*, 126(3):1447–1484.
- Shellman, S. M. (2001). Active learning in comparative politics: A mock german election and coalition-formation simulation. *PS: Political Science & Politics*, 34(4):827–834.
- Shweder, R. A., Much, N. C., Mahapatra, M., and Park, L. (2013). The” big three” of morality (autonomy, community and divinity). *Morality and Health*, page 119.

- Smith, A. (1822:1976). *The Theory of Moral Sentiments*. Liberty Fund.
- Smith, V. L. and Wilson, B. J. (2017). Sentiments, conduct, and trust in the laboratory. *Social Philosophy and Policy*, 34(1):25–55.
- Smith, V. L. and Wilson, B. J. (2019a). *Humanomics: Moral sentiments and the wealth of nations for the twenty-first century*. Cambridge University Press.
- Smith, V. L. and Wilson, B. J. (2019b). *Humanomics: Moral sentiments and the wealth of nations for the twenty-first century*. Cambridge University Press.
- Stanca, L. (2009). Measuring indirect reciprocity: Whose back do we scratch? *Journal of Economic Psychology*, 30(2):190–202.
- Steiger, E.-M. et al. (2014). See no evil: Information chains and reciprocity. *Journal of Public Economics*, 109:1–12.
- Sugden, R. (1998). Normative expectations: the simultaneous evolution of institutions and norms. In Ben-Ner, A. and Putterman, L., editors, *Economics, values and organization*, chapter 2, pages 73–100. Cambridge University Press, Cambridge, United Kingdom.
- Sutter, M., Haigner, S., and Kocher, M. G. (2010). Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. *The Review of Economic Studies*, 77(4):1540–1566.
- Tsang, J.-A. (2006). Gratitude and prosocial behaviour: An experimental test of gratitude. *Cognition & emotion*, 20(1):138–148.
- Tsang, J.-A. (2007). Gratitude for small and large favors: A behavioral test. *The Journal of Positive Psychology*, 2(3):157–167.
- Tversky, A. and Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.
- Ule, A., Schram, A., Riedl, A., and Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science*, 326(5960):1701–1704.
- van Apeldoorn, J. and Schram, A. (2016). Indirect reciprocity; a field experiment. *PloS one*, 11(4):e0152076.
- Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

- Wang, Z., Chen, T., and Wang, Y. (2017). Leadership by example promotes the emergence of cooperation in public goods game. *Chaos, Solitons & Fractals*, 101:100–105.
- Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(6):1063.
- Wedekind, C. and Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, 288(5467):850–852.
- Wilson, B. J., Jaworski, T., Schurter, K. E., and Smyth, A. (2012). The ecological and civil mainsprings of property: An experimental economic history of whalers' rules of capture. *The Journal of Law, Economics, and Organization*, 28(4):617–656.
- Winking, J. and Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, 34(4):288–293.
- Yang, Y., Onderstal, S., and Schram, A. (2016). Inequity aversion revisited. *Journal of Economic Psychology*, 54:1–16.
- Zizzo, D. J. (2013). Do dictator games measure altruism? In *Handbook on the Economics of Reciprocity and Social Enterprise*. Edward Elgar Publishing.