Belief Formation and Prosocial Behavior

Egon Tripodi

Thesis submitted for assessment with a view to obtaining the degree of Doctor of Economics of the European University Institute

Florence, 05 November 2020

European University Institute
Department of Economics

Belief Formation and Prosocial Behavior

Egon Tripodi

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Examining Board

Prof. David K. Levine, EUI, Supervisor
Prof. Michele Belot, Cornell University
Prof. Uri Gneezy, Rady School of Management, UC San Diego
Prof. Lorenz Götte, Institute for Applied Microeconomics, University of Bonn

European University Institute

DEPARTMENT OF ECONOMICS

Researcher declaration to accompany the submission of written work

I **Egon Tripodi** certify that I am the author of the work **Belief Formation and Prosocial Behavior** I have presented for examination for the PhD thesis at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the *Code of Ethics in Academic Research* issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. [quotation from it is permitted, provided that full acknowledgement is made.] This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party.

Statement of inclusion of previous work (if applicable):

I confirm that chapter 1 was jointly co-authored with Peter Schwardmann and Joel van der Weele and I contributed 33% of the work.

I confirm that chapter 2 was jointly co-authored with Lorenz Goette and I contributed 50% of the work.

I confirm that chapter 3 was jointly co-authored with Christian J. Meyer and I contributed 50% of the work.

Signature and Date:

October 16, 2020

ACKNOWLEDGEMENTS

I only made it through the PhD and the academic job market because very smart people invested their time to help me become a better economist, many have supported me when I was stuck or it was not my best day, and some have taught me to pause and enjoy the (sporadic) moments of professional accomplishment. I want to begin by thanking my partner Stephanie März for having done all these things.

I am extremely grateful to my advisor David Levine for helping me find my way of doing research. I was not the easiest student to supervise: I started the PhD thinking I could be interested in finance, only to start working on market design problems and wanting to use experimental methods to study economics, towards my third year. David was able to offer excellent guidance on any topic. I also benefited tremendously from the guidance of my second advisor Michele Belot, who took me on soon after she joined the EUI. I am so grateful for all her encouragement to pursue my most ambitious research ideas and for the expert guidance on how to execute them well. I owe so much to my wonderful co-authors Adam Altmejd, David Birke, Lorenz Goette, Felix Kölle, Christian Meyer, Peter Schwardmann, Simone Quercia, and Joel van der Weele. My research is so much better because of them. Special shout-out: to Lorenz for hosting me in Bonn for the longest research visit in modern history, for mentoring me, for watching me present my job market paper more times than I can count, and for teaching me the grit I thought I couldn't have; to Christian for being my partner in crime (ok, research is not a crime) since the beginning. Immense gratitude goes to the research support staff at the EUI and in Bonn. Especially, Simone Jost, Ilona Krupp, Sarah Simonsen, Jessica Spataro, Julia Valerio, Lucia Vigna. Thanks to all the friends I have made through (and with whom I've shared the pains of) grad school. Especially, Alessandro Ferrari, Anna Rogantini, Mathijs Janssen and Matthias Schmidtblaicher for being particularly present. Thanks also to my family.

A final word to Stefano Fenoaltea, who recently passed away. He inspired me early on to become an economist. He was a mentor and a friend. He will be dearly missed.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation consists of three self-contained essays on belief formation and on the role of beliefs for prosocial behavior.

The first chapter is co-authored with Peter Schwardmann and Jöel van der Weele. Does the wish to convince others lead people to persuade themselves about the factual and moral superiority of their position? We investigate this question in field experiments at two international debating competitions that randomly assign persuasion goals (pro or contra a motion) to debaters. We find evidence for self-persuasion in incentivized measures of factual beliefs, attitudes, and confidence in one's position. Self-persuasion occurs before the debate and remains after the debate. Our results lend support to interactionist accounts of cognition and suggest that the desire to persuade is an important driver of opinion formation.

The second chapter is co-authored with Lorenz Goette. We propose a novel experiment that prevents social learning, thus allowing us to disentangle the underlying mechanisms of social influence. Subjects observe their peer's incentives, but not their behavior. We find evidence of conformity: when individuals believe that incentives make others contribute more, they also increase their contributions. Conformity is driven by individuals who feel socially close to their peer. However, when incentives are not expected to raise their peer's contributions, participants reduce their own contributions. Our data is consistent with an erosion of norm-adherence when prosocial behavior of the social reference is driven by extrinsic motives, and cannot be explained by incentive inequality or altruistic crowding out. These findings show scope for social influence in settings with limited observability and offer insights into the mediators of conformity.

The third chapter is co-authored with Christian J. Meyer. We study incentivized voluntary contributions to charitable activities. Motivated by the market for blood donations in Germany, we consider a setting where different incentives coexist and agents can choose to donate without receiving monetary compensation. We use a

model that interacts image concerns of agents with intrinsic and extrinsic incentives to donate. Laboratory results show that a collection system where compensation can be turned down can improve the efficiency of collection. Image effects and incentive effects do not crowd each other out. A significant share of donors turn down compensation. Heterogeneity in treatment effects suggests gender-specific preferences over signaling.

Chapter 1

Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions

1.1 Introduction

How people form beliefs has been the subject of longstanding inquiry in the social sciences. Standard economic theory posits that agents interpret new evidence by using Bayes' rule, in a process of truth approximation. A large literature in behavioral economics proposes that people are boundedly rational and use heuristics in their attempts to discover the truth in complex information environments. Instead, an influential set of recent papers emphasizes the fundamentally social nature of human reasoning and belief formation that originates from the need to impress and persuade others (Mercier and Sperber, 2011; Von Hippel and Trivers, 2011; Kurzban, 2012; Mercier, 2016; Bénabou, Falk, and Tirole, 2019).

This "interactionist" approach maintains that our reasoning processes have developed to act more like a "press secretary" than a "scientist" (Kurzban, 2012). In the process of persuading others, we align our own beliefs and convictions with our political and economic goals.[1] This idea provides a unifying explanation for a number of well-documented cognitive phenomena in behavioral economics and social psychology. For instance, it explains how confirmation bias, partisanship and overconfidence arise from the wish to convince others of our opinions, politics and ability (Mercier and Sperber, 2011). However, despite its wide scope, there is no direct test of the interactionist approach in an ecologically valid setting. One key problem is that, in the field, the direction of causality between private views and the wish to persuade is usually unclear.

We confront this identification challenge in the context of two international debat-

[1]The interactionist approach therefore departs from the common assumption in economic models of communication that *senders*' beliefs are not systematically affected by their persuasion incentives (Milgrom, 1981; Crawford and Sobel, 1982; Kamenica and Gentzkow, 2011).

ing competitions where we investigate the causal effect of persuasion goals on the formation of beliefs and attitudes, a phenomenon we call "self-persuasion".[2] The debating competitions take place in Munich and Rotterdam and attract members from debating clubs from all over Europe. Across several rounds, participants debate motions on topical political issues such as freedom of movement in the European Union, the merits of geoengineering, the appropriate power of trade unions, and the regulation of big technology companies. In this context, we elicit beliefs and attitudes surrounding the debated motions in each of the qualifying rounds of the tournament, both before and after the debates. To make sure that our elicitations reflect true beliefs and attitudes, we incentivize reports with an incentive compatible scoring rule.

Several features of debating tournaments make them ideally suited for testing the interactionist approach. First, debaters are randomly assigned to positions pro or contra the motion shortly before the start of the debate. This allows us to make causal inferences about the effect of persuasion goals. The nature of the randomization solves two problems that may arise in the identification of self-persuasion. Because the assignment is randomized explicitly, participants know not to infer anything about the merit of the assigned debating position—a problem with many experimental designs used to study politically motivated reasoning (Tappin, Pennycook, and Rand, 2019). Moreover, since the randomization is a natural aspect of the tournament, participants do not view it as experimental variation, ameliorating concerns of potential experimenter demand effects. Another unique aspect of our setting is that debaters' intrinsic motivation to be persuasive is high. A panel of experienced judges evaluates the quality of each debater's arguments, determining his or her success in the tournament and subsequent status in the debating community. These incentives for persuasion mimic those of professionals in politics and law. It is no coincidence that many famous politicians and lawyers honed their skills by taking part in competitive debating.[3]

[2]Using terminology from the persuasion literature (DellaVigna and Gentzkow, 2010), a persuasion goal is a behavior (e.g. *vote for A*) or a view (e.g. *A is a good policy*) that a sender wants a receiver to take.

[3]For instance, prominent Brexiteers Boris Johnson and Michael Gove were president of the Oxford Union, a renowned debate club. Other prominent politicians who were part of debating societies include Nancy Pelosi, Jimmy Carter, Margaret Thatcher and John Major. See either the site of the National Speech and Debate Association or this site for partial lists of famous former debaters.

We find strong evidence for self-persuasion in debaters' beliefs and attitudes, measured after persuasion goals are assigned but before the debate begins. First, participants are more likely to believe that a factual statement is true if the statement strengthens an argument supporting their position. Second, in a monetary allocation task between charities, debaters shift donations towards goal-aligned charities. Third, debaters become more confident about the strength of the arguments on their side of the motion, as measured by the subjective probability that teams arguing the same side of the motion in other debates will win. For all three outcomes, self-persuasion is measured as the gap in beliefs or attitudes between debaters arguing against and those arguing in favor of a motion. Importantly, beliefs elicited before the assignment of persuasion goals confirm that there are no pre-treatment differences between these two groups.

We also investigate whether the debate itself mitigates the effect of self-persuasion by exposing participants to arguments from the other side. We find weak convergence in beliefs and none for attitudes, so polarization in both measures persists after the debate. As a result, debaters leave the tournament more polarized than they started. Since debaters are never asked the same question twice, the persistence of polarization is not driven by concerns for consistency. Furthermore, we find that self-persuasion effects are at least as strong as persuasion by the winning arguments in the debate. Thus, at least in our setting, self-persuasion causes the exchange of ideas to be a catalyst of polarization rather than an antidote to it.

Our findings lend support to an interactionist account of human cognition in which persuasion goals drive non-Bayesian belief and attitude formation (Mercier and Sperber, 2011; Von Hippel and Trivers, 2011). Our data also allow us to comment on the mechanisms underlying self-persuasion. Mercier and Sperber, 2011 argue that self-persuasion is a *by-product* of persuasion, resulting from a cognitive failure to account for our disproportionate investment in finding the strengths in our own and the weaknesses in our interlocutor's position. Instead, Von Hippel and Trivers, 2011 argue that self-persuasion is *strategic*: people self-deceive because believing in the moral and fac-

tual superiority of their position makes them more persuasive. To investigate these channels, we ask debaters how many arguments they generated for each position during their preparation time. We find that arguments are highly skewed towards their own position, and that this imbalance can explain about *half* of the treatment effect. Thus, our data suggests that self-persuasion is partly driven by debaters' naive overappreciation of their own biasedly generated arguments and partly driven by other mental processes likely due to self-deception.

Previous experimental research from the laboratory provides evidence consistent with self-persuasion. Classical cognitive dissonance experiments from social psychology have demonstrated that "forcing" people to make counter-attitudinal statements affects subsequent stated attitudes (e.g. Festinger and Carlsmith, 1959; Elliot and Devine, 1994). In economic experiments, subjects self-deceive to justify giving self-serving financial advice (Chen and Gesche, 2017; Gneezy et al., 2020), and to more effectively persuade other subjects in the experiment (Smith, Trivers, and Hippel, 2017; Schwardmann and Weele, 2019; Solda et al., 2019). Studies on "role-induced bias" investigate the effect of random role assignment on beliefs or attitudes. For instance, Janis and King, 1954 show that role-playing that involves overt verbalization of arguments affects opinions. O'Neill and Levings, 1979 find that experimental subjects selectively scan evidence for arguments in favor of a position they were asked to argue. A number of studies randomly assign subjects to the role of plaintiff or defendant in courtroom simulations (see Engel and Glöckner, 2013, for evidence and a review). Most prominently, Babcock et al., 1995 show that the assigned role leads subjects to change their fairness judgements and their assessment of an actual judge's verdict.

Our paper provides the first *field* evidence that persuasion goals drive non-Bayesian belief and attitude formation. Our unique setting allows us to establish that the phenomenon is not an artifact of the laboratory protocols in previous experiments. Moreover, the fact that our results obtain at prestigious international tournaments and in subjects with years of debating experience demonstrates that competitive incentives and experience do not lead people to conform to "neo-classical" assumptions (List,

2003). We are also able to provide a broader picture than previous studies, which typically focus on a limited, context-dependent set of outcomes, like perceptions of a selected legal case measured at a single point in time. We show that self-persuasion affects confidence, factual beliefs and affective attitudes both before *and* after the debate, measured across a range of politically relevant topics and questions, using incentives for accuracy for all variables. Finally, the clean manipulation of persuasion goals allows us to speak to prominent theories about the nature of human reasoning and social influences in belief and attitude formation, which have seen relatively little testing so far.

More broadly, we contribute to a literature on motivated cognition that investigates how affective and functional goals influence belief formation (see Kunda 1990; Bénabou and Tirole 2016; Gino, Norton, and Weber 2016 for surveys). This literature has seen almost no testing of the effect of goals on beliefs in ecologically valid settings. Two exceptions are Di Tella, Galiant, and Schargrodsky, 2007, who find that squatters attitudes toward markets become more favorable after being granted legal titles to their land, and Oster, Shoulson, and Dorsey, 2013, who show that people at risk of Huntington disease are prone to wishful thinking. Our study differs from these by investigating the effect of persuasion goals. Furthermore, we consider a more tightly controlled setting that, unlike previous field studies, allows for the incentivized elicitation of beliefs and attitudes (Schlag, Tremewan, and Weele, 2015).

There is also an immediate connection of our results with the empirical literature on polarization and political opinion formation. Researchers across the social sciences have used laboratory experiments to show how confirmation bias and selective parsing of arguments can lead to attitude polarization (Lord, Ross, and Lepper, 1979; Sunstein, 2002). Several different mechanisms have been proposed to fit these data (Taber and Lodge, 2006; Kahan, 2015; Fryer, Harms, and Jackson, 2018).[4] We show that per-

[4]Within economics, some theories have extended standard Bayesian belief updating to better capture the role of social interactions in belief formation. These papers formalize drivers of polarization that work through the identification with social groups (Gennaioli and Tabellini, 2019) as well as the production of narratives to interpret historical data (Eliaz and Spiegler, 2018) or to influence the behavior of others (Bénabou, Falk, and Tirole, 2019; Foerster and Weele, 2018).

suasion motives induce polarization on a range of cognitive and non-cognitive measures, suggesting that a number of different mental processes are at work. Here, using incentives for truthful reporting is crucial, as Bullock et al., 2015 show that voters display up to 80 percent less polarized attitudes when their answers are incentivized for accuracy.

Relatedly, our analysis of competitive debating contributes to the discussion about the merits of deliberative democracy. According to the ideal of deliberative democracy the exchange of opinions helps to resolve conflicts and foster social consensus (e.g. Habermas, 1984; Elster, 1998; Gutmann and Thompson, 2004). By contrast, the literature on polarization has shown that deliberation can have exactly the opposite effect (Kuhn, Shaw, and Felton, 1997), and promote radicalization in interactions between like-minded people (Sunstein, 2002). The conditions for deliberation to work best are a matter of active debate in political science (e.g. Thompson, 2008; Mercier and Landemore, 2012). Our results show that in a setting where individuals' chief motivation is to prevail over their competitors, even the *prospect* of debate increases polarization and that the subsequent debating does little to decrease it.

The remainder of the paper is structured as follows. Section 1.2 describes the setting, sample, and procedures of the field experiment. Section 1.3 presents results on the effects of persuasion goals on privately held views, and on the effects of debating on polarization. Section 1.4 provides evidence to inform a discussion on the psychological mechanisms of self-persuasion and the relation between self-persuasion and debater success. Section 1.5 concludes by discussing some implications of our results.

1.2 Experimental Setting

Competitive debating is popular among high-school and university students. Many universities have debating societies that organize local or international tournaments, the most prestigious of which include the North American, European and World Championships. Motions relate to topical issues in politics such as immigration, climate change and the regulation of new technology. In contrast to debates between experts

or politicians, competitive debaters are randomly assigned to defend particular positions, which may or may not correspond to their private opinions.

Our study took place at two international debating competitions in March 2019: the *Munich Research Open*, and the *Erasmus Rotterdam Open*. Both tournaments followed the British Parliamentary (BP) debating format, in which debates take place with two teams of two debaters arguing in favor of (Proposition) and two teams against (Opposition) a given motion. Persuasion goals (Proposition/Opposition) are randomly assigned to teams and all speakers have equal time to present their arguments. The motions are prepared by chief adjudicators before the tournament, and revealed to the debating teams fifteen minutes ahead of the debate. They are designed such that there are valid arguments for both sides. Debaters are evaluated on the quality of their arguments by a panel of three expert judges, who themselves have experience as debaters.

The competitions featured 52 (Munich) and 48 (Rotterdam) teams and took place in two phases. In the preliminary phase of the tournament (*in-rounds*), all teams debate multiple times: each round features a motion that all teams debate in parallel sessions. In each round, teams are partitioned into 13 (Munich) or 12 (Rotterdam) parallel debating sessions of four teams each using a conditional random assignment. Teams accumulate points that depend on their evaluation and determine who advances to the knock-out phase of the competition. Appendix A.1 provides further details on the BP debating format.

1.2.1 Sample

Participants of international debating competitions in the BP format are predominantly undergraduate and graduate students, who are members of debating societies. They accumulate debating experience through tournament participation and regular meetings at the debating societies of their university, and sometimes also from a high-school debating career. The characteristics of BP debating attracts speakers with strong

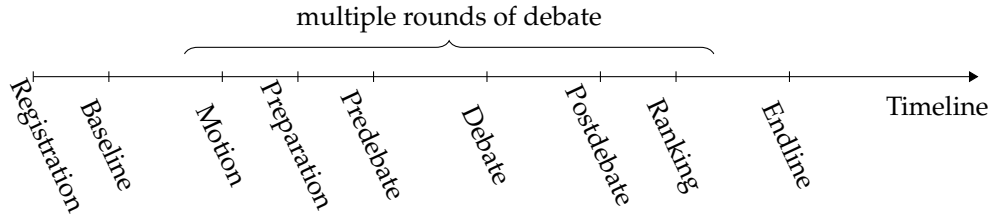
analytical skills, fast thinking and a breadth of knowledge.[5]

On average, our sample has spent more than two years in debating, has qualified for more than three semi-finals of an international tournament, is about 22 years old, and tends to hold a relatively liberal ideology. Men are somewhat over-represented and the sample is very international – less than 25 percent of participants hold nationality from the country where the tournament is hosted. The sample is similar across the two tournaments in terms of age, local representation, political views, and time spent in debating. However, there are some differences in terms of the gender balance and past achievements: the share of female debaters is 17 percentage points higher in Munich than in Rotterdam, and debaters in Rotterdam have reached semi-finals in large international competitions more than twice as many times than debaters in Munich.

More importantly for the internal validity of our findings, in Table A.5 we show balance of individual characteristics and baseline views on topics related to each motion across debaters with different persuasion goals. For some of the questions we randomized the order across subgroups. In Table A.6 we show that individual characteristics are balanced also across these subgroups.

1.2.2 Research Design

We only collected data during the preliminary rounds of the competitions (five in Munich and four in Rotterdam) to maintain a balanced panel of observations. Debaters answered four main surveys with the following timing:



[5]Further discussion of the characteristics of debaters that take part in this format on the website of the American Parliamentary Debate Association.

1. **Baseline.** Administered at the very beginning of the tournament. Contains background questions as well as instructions on the quadratic scoring rule (QSR) – the procedure that we use throughout all surveys to elicit beliefs in an incentive compatible manner.

2. **Predebate.** Administered right after the preparation time of each debating session, just before the debate begins.

3. **Postdebate.** Administered right after each debate ends.

4. **Endline.** Administered after the fifth and last debate of the preliminary phase (Munich) or after the fourth round of the preliminary phase (Rotterdam).[6]

Our main survey measures are the following:

- **Factual beliefs.** These were factual statements that related to the motion, and debaters had to predict whether the statements were true or false. Factual statements were constructed such that, if they were true, one side of the debate would find them "convenient" in support of their arguments. We elicit Factual beliefs related to the motions at Baseline, Predebate, and Postdebate.

- **Attitudes:** We asked debaters to allocate money between a "neutral" charity and a charity that was aligned with one side of the motion. Each charity was described to respondents in a short paragraph on the same survey sheet. We elicit Attitudes related to the motions at Predebate, and Postdebate.

- **Confidence in proposition:** We elicited the subjective probability that a majority of parallel debates (excluding the debater's own debate) in the round will be won by the proposition side of the debate. This is a measure of the perceived advantage of a persuasion goal, independent of a speaker's confidence in her own ability. We elicit Confidence in proposition only at Predebate.

[6]This difference is due to different schedules of the tournaments. In both cases, the endline survey took place after the last round of a four-round day. In Rotterdam, the tournament started in the morning and had a full day with four rounds of debate. In Munich, the tournament started in the late afternoon with one round of debate and had four rounds of debate the day after.

Next we provide an example of a motion and an associated factual statement, charity and confidence question from the surveys. Section A.2 provides detailed examples of factual belief elicitations from motions in our debates.

Example of motion: This house regrets the EU's introduction of the freedom of movement.

Factual statement: More than 35% of UK citizens interviewed for the Eurobarometer in 2018 think that the Schengen Area has more disadvantages than advantages for the UK.

Charity: ACT4FreeMovement stands for Advocacy, Complaints, Trainings for Freedom of Movement. The organization campaigns for freedom of movement with EU citizens. The goal is to increase the capacity of EU citizens to effectively secure access to and knowledge of their rights, as well as build public awareness and political support for mobile citizen rights.

Confidence statement: Excluding the debate happening in this room, in at least half of the parallel debates of this round, one of the two teams on the Government side of this motion will rank 1st.

We incentivized our main outcome variables as follows. For the Factual beliefs and the Confidence elicitation, subjects were incentivized with a binarized quadratic scoring rule that paid in lottery tickets. By providing a report $r \in [0, 100]$, given the objective binary answer $R \in \{0, 1\}$, a subject receives a lottery ticket that paid off a monetary prize of 30 euros with the following winning probability

$$w = 1 - \left(R - \frac{r}{100}\right)^2.$$

Of all elicitations of this kind, only one was randomly selected to be paid at the end of the study. Our general instructions used both the mathematical equation, a simple quantitative illustration, and an intuitive explanation that incentives were designed so that the truthful reporting optimizes the likelihood of winning the prize of 30 euro (see Section A.9).[7]

[7] In theory, this procedure makes the quadratic scoring rule incentive compatible for all risk preferences (Hossain and Okui, 2013; Schlag and Van der Weele, 2013). Whether this is actually the case in practice is a matter of ongoing debate.

For the Attitude variable, subjects allocated up to 10 euro between two different charities, where the budget constraint was concave in order to discourage extreme choices. One of the choices was randomly selected and the experimenters made the charitable payments on the subjects' behalf.

In addition to these incentivized measures, we elicited some background variables, including gender, debating experience and performance, as well as some basic socio-demographics.[8] In our Endline survey, we also asked several questions on "impressions", for example, about factual statements and the goal of the research. These variables served to check the robustness of our main results. Table 1.1 summarizes how survey elements were distributed across the different surveys.

Debates were moderated by a panel composed of three (sometimes two) judges. These were experienced debaters themselves trained to evaluate debaters' speeches according to standardized international criteria. After the debate, judges deliberated in private to produce the "ballot", an official score sheet that consists of the technical score on the quality of arguments made by each debater in each debate and determines the ranking of teams in each debate. In addition, we asked judges to independently fill out a "judge survey" where they assign a broad persuasiveness score to each debater. We told judges that this score should consider quality of arguments as well as body language, tone, and other markers that make a speech persuasive to a general population.

Table 1.1: Content of Debater Surveys and Timing

			Incentivized Outcome Variables			
Survey	Timing	Background Info	*Factual beliefs*	*Attitudes (charities)*	*Confidence in proposition*	Impressions
Baseline	Beginning of tournament	X	X			
Predebate	Right before each debate		X	X	X	
Postdebate	After each debate		X	X		
Endline	After last debate					X

[8]The Baseline survey also included some incentivized factual knowledge "decoy" questions about topics not related to the motions. These questions served to obfuscate the elicitation of Factual Beliefs related to the motions and not give away the topics of the motions that were still secret at that point.

The four debater surveys as well as the judge survey were administered by an enumerator, who also attended the debate and filled out a separate "enumerator survey" that was designed to capture both objective and subjective measures of how heated debates were, and whether facts and charities included in the survey questions were mentioned by debaters to make their case. Enumerators were asked to take note of any anomaly that might have occurred during the debate.

The full content of all surveys is described in detail in Section A.9. Section A.10 provides all motions, survey questions and charities used for the attitude elicitation.

1.2.3 Survey Versions and Administration Procedures

Before each tournament, we interacted with the chief adjudicators to converge on a final set of motions for the debate. For each motion, we developed four factual questions (A, B, C, D) and found two motion-related charities (E, F). We varied the order in which factual questions and charities were presented between two different subgroups, as illustrated in Table 1.2. We created these subgroups in advance using lists of registered participants and identified a debater's subgroup by adding an ID number to their name tag.

The use of multiple questions in different orders assures that no debater answers the same question twice and that no result depends on the answer to a single question or the order in which questions were asked. It also eliminates the desire to provide consistent answers to repeated questions and reduces potential experimenter demand effects. Moreover, since baseline and predebate questions were different both within and across subgroups, participants could not be influenced through discussion of the answers with others.

The baseline survey was administered in a large common room after some introductory remarks by the organizers and one of the researchers. In this room, debaters were given 10 minutes to read carefully a set of general instructions for the surveys, and subsequently had 25 minutes to answer the baseline survey. The survey is similar for all participants except for the factual questions that directly relate to the in-rounds

motions, which differed between subgroups as displayed in Table 1.2.

Table 1.2: Distribution of Factual Questions and Charities Over Surveys

	Motion factual questions			Motion charities	
	Baseline	Predebate	Postdebate	Predebate	Postdebate
Subgroup 1	A	D	B, C	E	F
Subgroup 2	B	C	A, D	F	E

Note: Distribution of four factual questions per motion and two motion-related charities over surveys. Each letter corresponds to one factual question/charity.

In each debating round, the motions were announced in the central meeting room, and debaters made their way to the assigned debating room after announcements. Enumerators distributed the predebate survey in the separate debating rooms. While seated at their desks, debaters were given up to five minutes to answer and enumerators ensured that they did not use this time to prepare for the debate. At the beginning of the debate enumerators also distributed the judge survey, in which judges indicated their evaluations of persuasiveness. Judges had the entire debate session plus their regular judge deliberation time to fill out this survey.

After the predebate survey, the judges opened the debate. During the debate itself, which lasts about an hour, enumerators filled in their own surveys, noting down participant IDs and debate impressions. Once the judges declared the end of the debate, enumerators distributed the postdebate survey, which debaters had five minutes to answer.

The endline survey was administered just outside of each debate room right after the end of the last round of debates covered by our intervention. Debaters had twenty minutes to answer this survey, which they did in the corridors outside the debating room. Enumerators insisted with subjects to not interact with others or mobile devices during this time.

1.3 Results

Our main focus lies on the question of how persuasion goals affect self-persuasion, as measured by our predebate elicitations on Factual Beliefs, Attitudes and Confidence. A secondary question relates to the role of the debate itself in shaping the divergence of views among debaters.

1.3.1 Self-Persuasion

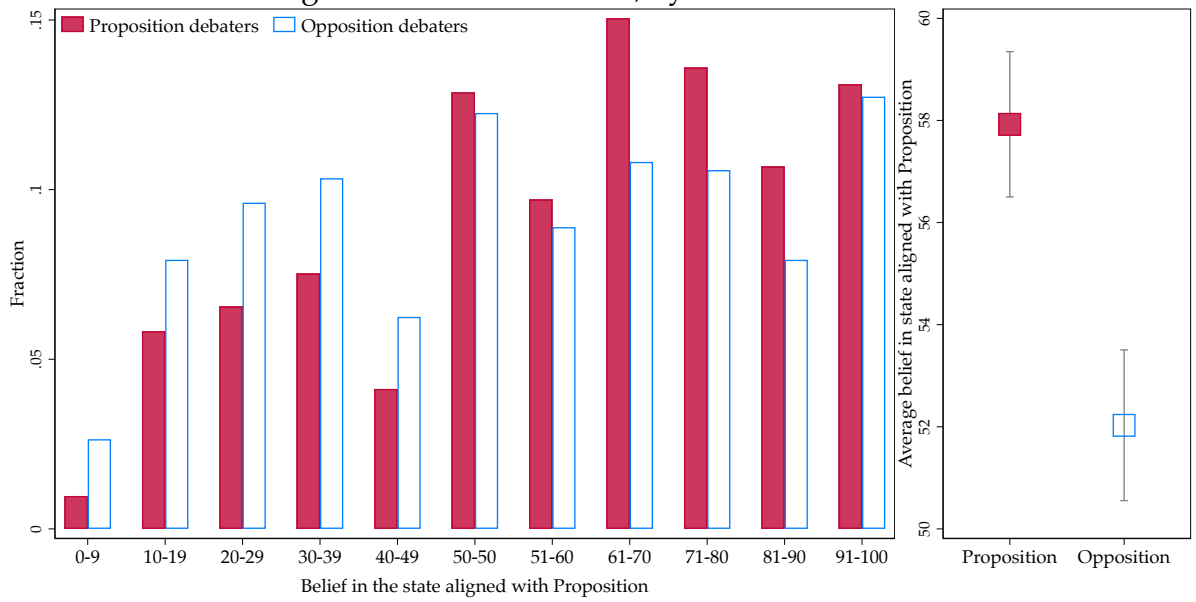
We compare differences in our main outcome variables, i.e. Factual Beliefs, Attitudes and Confidence, between debaters on the proposition and on the opposition side of the motion. We look at the predebate elicitations, which reflect only the cognitive processes taking place in the 15 minute preparation period after persuasion goals are assigned, and are not affected by the actual debating itself. In Table A.8 we evaluate gender differences in these effects. In Section A.7, we assess the extent to which debaters were able to infer our research hypotheses, and use this as input to a robustness analysis of our self-persuasion results to experimenter demand effects.

Do Persuasion Goals Affect Factual Beliefs?

For every factual belief question, one state (e.g. the statement is true) is more favorable to the proposition of the motion than the other state (e.g. the statement is not true). In order to compare questions, we transform each belief into the subjective probability that the state that favors the proposition is true. When a factual statement is favorable to the proposition (opposition), this corresponds to the reported subjective probability that the statement is true for speakers on the proposition (opposition) side of the debate, and to the complementary probability for speakers on the opposition (proposition) side. More background information on which states are considered favorable to the proposition is provided in Section A.2.

Figure 1.1 reports beliefs in the state aligned with the proposition. The left panel shows a histogram of beliefs grouped in equally spaced probability brackets, except for the intermediate 50-50 category. The right panel shows that the difference in av-

Figure 1.1: Factual Beliefs, by Persuasion Goal



Note: Predebate beliefs elicited from debaters over multiple rounds are pooled and each report $r \in [0, 100]$ is transformed as the complement to 100 if the report is not aligned with the proposition. In the left panel, the pooled and transformed beliefs are grouped in equally spaced probability brackets – except for the intermediate 50-50 category. In the right panel, we report averages of this outcome by position in the debate, and ranges indicate standard errors.

erage beliefs between the two groups is 5 percentage points. These data show that debaters are more likely to believe in the answer that favors the proposition, if they themselves are in the proposition.[9]

To assess the statistical significance and the magnitude of this effect, and gain greater comparability of subjective probabilities on the truthfulness of different factual statements, we conduct both a normal standardization of the reported belief (separately for each question) and adjust the sign of the standardized belief. In turn, a positive (negative) sign of such standardized outcome captures alignment with the state that favors the proposition (opposition). After adjusting the sign, the standardized belief remains normally distributed with zero mean and unit standard deviation. This transformation yields an individual level outcome variable $b_{i,m}$ that admits a straightforward interpretation in terms of debater i's belief alignment with the proposition of motion m.

[9]Note that on both sides of the debate, debaters are more likely to believe that the answer favors the proposition. This is partly driven by the correct answer being aligned with the proposition relatively more frequently.

We estimate the gap in belief alignment with the proposition in a regression model

$$b_{i,m} = \alpha_i + \beta Proposition_{i,m} + \delta_m + \varepsilon_{i,m} \tag{1.1}$$

in which we include motion fixed effects δ_m and debater fixed effects α_i and allow for the error term to be correlated within each team of debaters.

Table 1.3 shows the results of the estimation. We confirm the finding that proposition debaters report beliefs that are markedly different from the beliefs reported by opposition debaters. Because of the randomized allocation of persuasion goals, this pattern cannot be explained by pre-existing differences between debaters on the two sides of the debate and has a causal interpretation. The lack of pre-existing differences in prior beliefs (Table A.5) highlights that a shift in beliefs depending on the assigned position violates the basic Martingale property of Bayesian beliefs—that posteriors in expectation equal priors. Factual Beliefs of proposition debaters are 21.5 percent of a standard deviation (column 1, $p < 0.001$) closer to the proposition alignment. This effect is robust to the omission of fixed effects (column 2) and the inclusion of controls (column 3). The magnitude of this effect is also meaningful in terms of monetary losses: debaters make average expected earnings of 27.4 euro (24.7 euro) on factual belief elicitations for which the correct answer is aligned with (against) their persuasion goal.

Table 1.3: Panel Regressions for Effects of Persuasion Goals on Factual Beliefs

	Beliefs align with proposition		
	(1)	(2)	(3)
Debater in proposition	0.215***	0.217***	0.203***
	(0.062)	(0.061)	(0.062)
Socio-demographic and experience controls			✓
Debater fixed effects	✓		
Round fixed effects	✓	✓	✓
Observations	884	884	851

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level. Socio-demographic controls include age, gender, and an indicator for whether the debater's nationality is from the country that hosts the competition. Experience controls include the reported number of international tournaments in which the debater has made it to semi-finals, and a categorical variable capturing the number of years the debater has been actively debating. Some observations are lost in column (3) due to missing control variables.

Result 1 (Factual Self-Persuasion)**.** *Persuasion goals make individuals distort their perception of facts in the direction that strengthens the positions they need to defend.*

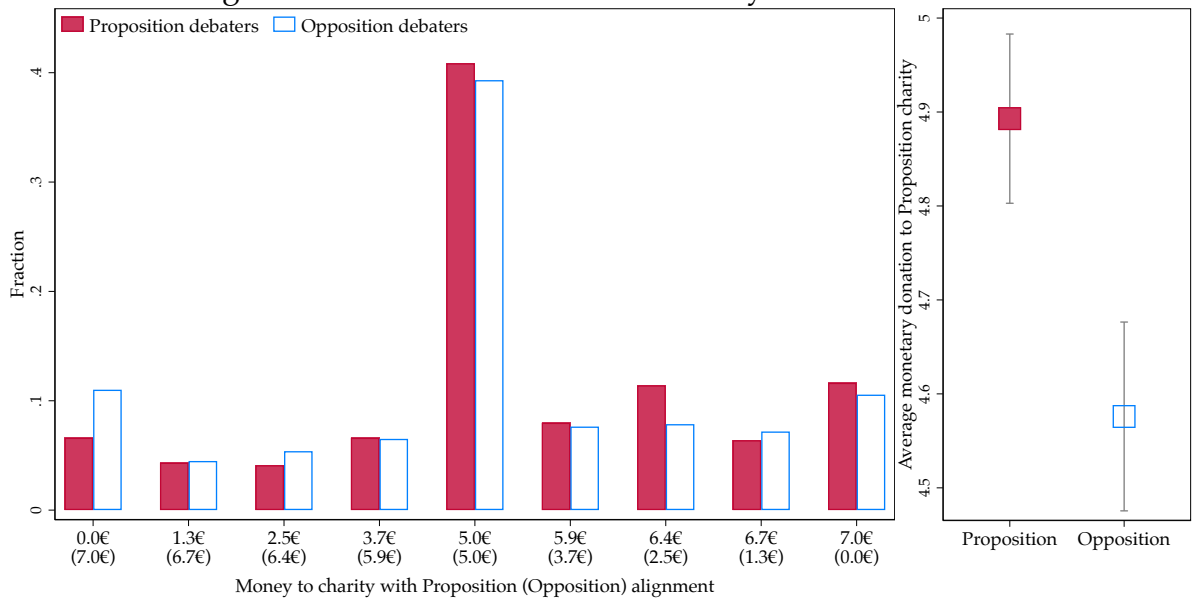
Do Persuasion Goals Affect Attitudes?

We measure attitudes towards the persuasion goal by how much money the debater allocates to a charitable cause that is aligned to her persuasion goal relative to a neutral charity. Remember that allocations lie on a concave budget constraint to encourage choices in the interior of the donation space.

The left panel of Figure 1.2 depicts donation choices across all motions. Allocations on the right side favor the charity aligned with the proposition and choices on the left side favor the charity aligned with the opposition. About 40 percent of allocation choices feature an equal split. Among the remaining observations we see a tendency for debaters to favor charities that are aligned with their persuasion goal. The right panel of Figure 1.2 shows that debaters in the Proposition role pay on average about

31 eurocents more to the Proposition charity.

Figure 1.2: Chosen Donation Bundles by Persuasion Goal



Note: Predebate allocations of charitable donations over multiple rounds are pooled. In the left panel each allocation $a \in \{0, .., 8\}$ is transformed as the complement to 8 if the allocation does not favor the charity with relative proposition alignment. In the right panel, we report average monetary donations to the charity with Proposition alignment by position in the debate, and ranges indicate standard errors.

To estimate the size and statistical significance of the effect, we use a fixed effects regression framework similar to model 1.1, in which the ordinal outcome capturing how favorable the debater's allocation is to the proposition charity is treated as a continuous variable.[10] We complement this analysis with regressions that use as continuous outcomes directly the monetary amounts donated to proposition and opposition charities implied by the bundle chosen by the debater.

Table 1.4 presents the results of the estimation. We confirm the impressions from visual inspection of the pooled outcomes: persuasion goals lead proposition debaters to choose an allocation of charitable donations that is 0.306 positions more favorable to the charity with proposition alignment (column 1, $p = 0.023$).[11] Columns (4) and (5)

[10]The more appropriate regression model would take into account the discrete ordinal nature of the outcome variable. However, ordered log-odds estimated from ordered Logit models are very hard to interpret. We provide panel estimates of the ordered Logit model in Table A.7. These are qualitatively very similar and support the main analysis presented here.

[11]Table A.8 shows that this result is more pronounced for men than it is for women.

aid the interpretation of this point estimate: From a total concave budget to allocate between two charities that can range from 7 to 10 euro, proposition debaters tend to sacrifice 0.239 euro that could go to the charity with opposition alignment to give 0.316 euro more to the charity with proposition alignment. While the magnitudes of these effects in monetary terms may seem small, one should keep in mind that concavity of the budget is such that efficiency seeking preferences attract subjects toward the intermediate allocation and make self-persuasion harder to detect. The asymmetry of this transfer is largely due to the frequency of extreme aligned allocations among opposition debaters.

Table 1.4: Panel Regressions for Effect of Persuasion Goals on Attitudes

	Donation bundle favorable to Proposition charity			Money to charity in Proposition	Money to charity in Opposition
	(1)	(2)	(3)	(4)	(5)
Debater in proposition	0.306**	0.297**	0.300**	0.316***	-0.239*
	(0.132)	(0.136)	(0.145)	(0.122)	(0.124)
Socio-demographic and experience controls			✓		
Debater fixed effects	✓			✓	✓
Round fixed effects	✓	✓	✓	✓	✓
Observations	883	883	850	883	883

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level. Socio-demographic controls include age, gender, and an indicator for whether the debater's nationality is from the country that hosts the competition. Experience controls include the reported number of international tournaments in which the debater has made it to semi-finals, and a categorical variable capturing the number of years the debater has been actively debating. Some observations are lost in column (3) due to missing control variables.

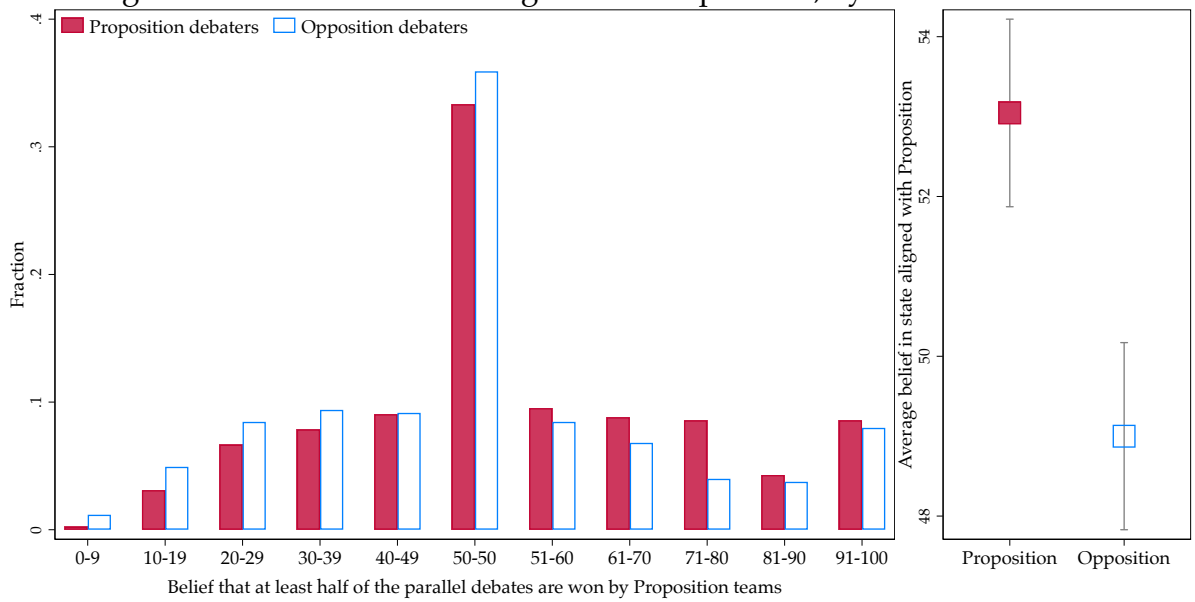
Result 2 (Moral Self-Persuasion)**.** *Individuals favor social causes aligned with their persuasion goals.*

Do Persuasion Goals Affect Confidence in One's Side of the Debate?

Our third outcome measure is debaters' Confidence in the strength of the proposition side of the debate. This is reported by debaters as the probabilistic prediction that at least half of the parallel debates will be won by proposition teams. Importantly, since debaters are betting on the outcome of the parallel debates and not on their own performance, this belief reflects the perceived strength of the debating position abstracting from beliefs in their own ability.

Figure 1.3 depicts probabilistic beliefs that the proposition will win in more than half of the parallel sessions. The left panel shows a histogram of beliefs grouped by equally spaced probability brackets – except for the intermediate 50-50 category. Beliefs are polarized across the two sides of the debate: 38 percent of the beliefs reported by proposition debaters lie above 50 percent, while only 30 percent of opposition debaters state beliefs higher than 50 percent. The right panel shows that difference in average beliefs between the two groups.

Figure 1.3: Perceived Advantage of the Proposition, by Persuasion Goal



Note: Predebate Confidence in the proposition, measured as the probability that at least half of the parallel debates are won by proposition teams, reported from debaters over multiple rounds are pooled. In the left panel, the pooled confidence reports are then grouped in equally spaced probability brackets – except for the intermediate 50-50 category. In the right panel, we report averages of this outcome by position in the debate, and ranges indicate standard errors.

When it comes to the empirical distribution, the proposition team wins the majority of parallel debates in each round only 43 percent of the time. Debaters' average probabilistic beliefs in this event are 49 percent in the opposition and 53 percent in the proposition. Hence, all debaters tend to overestimate the chances of proposition teams in these debates, but debaters in the proposition exhibit a greater bias.

Table 1.5: Panel Regressions for Effects of Persuasion Goals on Confidence

	Confidence in proposition teams		
	(1)	(2)	(3)
Debater in proposition	4.531***	4.389***	4.319***
	(1.498)	(1.492)	(1.554)
Socio-demographic and experience controls			✓
Debater fixed effects	✓		
Round fixed effects	✓	✓	✓
Observations	883	883	850

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level. Socio-demographic controls include age, gender, and an indicator for whether the debater's nationality is from the country that hosts the competition. Experience controls include the reported number of international tournaments in which the debater has made it to semi-finals, and a categorical variable capturing the number of years the debater has been actively debating.

To estimate the effects of persuasion goals on the perceived strength of the proposition, we can directly use the raw belief data on Confidence in the proposition as outcome in a fixed effects regression framework similar to equation (1.1).[12] The results of this analysis are reported in Table 1.5. Debaters in proposition teams are significantly more likely to believe that proposition teams will win the majority of debates. The reported probability assigned to the event that the majority of parallel debates will be won by proposition teams is higher by about 4.5 percentage points (column 1,

[12] An ordered Logit random effects model could be estimated instead to account for the strong discontinuity of the distribution of the outcome at 50-50. The estimates from that model are qualitatively identical to the ones presented in Table 1.5.

$p < 0.005$) for debaters who propose the motion relative to those who oppose it. This estimated effect is also about 20 percent of a standard deviation in the outcome – a similar magnitude to the self-persuasion effects on factual beliefs reported in the previous section, and also remarkably similar to estimates in Schwardmann and Weele (2019).

Result 3 (Confidence)**.** *Persuasion goals make individuals relatively more confident about the strength of the positions they defend.*

1.3.2 Debates and the Dynamics of Disagreement

Deliberative democracy depends on the power of debate to moderate disagreement and move opinions towards the side with the stronger arguments. In this section, we investigate whether debates fulfil these expectations. To this end, we compare Beliefs and Attitudes at the start of the debate, as measured in the predebate survey, with those at the end, as expressed in the postdebate survey. We first look at how the debate affects the polarization induced by self-persuasion. We then investigate the evidence for persuasion by the winning arguments in the debate.

Dynamics of Polarization

As a measure of polarization we use the sample variance σ^2 in beliefs and attitudes. To track disagreement both within and between the proposition and opposition sides, we decompose this variance in *between* group and *within* group variation. In particular, σ^2 can be written as the weighted average of Mean Squares Between groups (MSB) and Mean Squares Within groups (MSW) as follows[13]

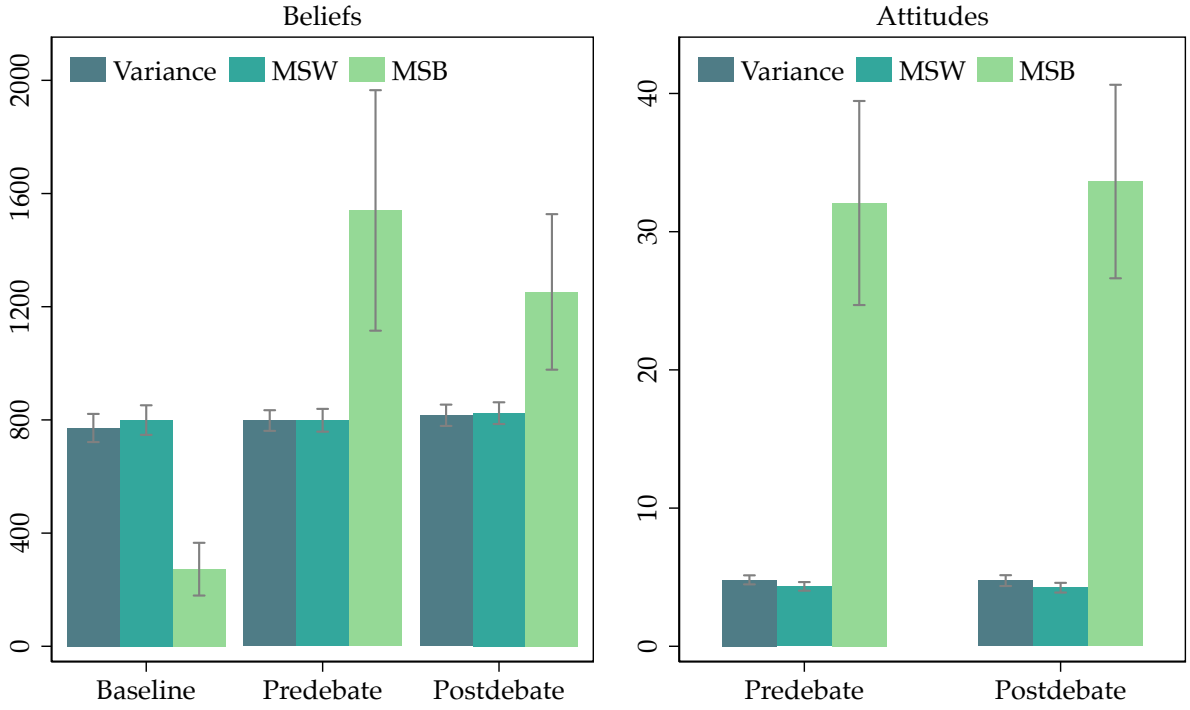
$$\sigma^2 = \frac{k-1}{n} MSB + \frac{n-k}{n} MSW,$$

where n is the sample size and k is the number of groups. For each Factual Belief and Attitude elicitation have two subgroups ($k = 2$) and a sample of about $n = 50$

[13]Using the well known decomposition of the Total Sum of Squares in the sum of Between Sum of Squares (BSS) and Within Sum of Squares (WSS), and the definition of mean squares as the sum of these squares statistics over their degrees of freedom ($MSB := BSS/(k-1)$, and $MSW := WSS/(n-k)$).

observations (this represents half of the participants in each tournament, as we randomized the order of elicitations between two subgroups). We have two questions and two charities for each of the nine different motions, leading to 18 observations of within and between group polarization for each variable. This allows us to statistically compare the distributions of Total variance (σ^2), MSB and MSW across different stages of the debate.

Figure 1.4: Variance Decomposition of Beliefs and Attitudes



Note: For each elicitation of factual beliefs and attitudes from an identical question that debaters answer in the same survey we have a sample of about 50 responses from both proposition and opposition debaters. Over both tournaments we have 18 belief questions elicited at baseline and postdebate, 18 belief questions elicited at predebate and postdebate, and 18 allocations of donations between different charities elicited at predebate and postdebate. Ranges indicate standard errors.

Figure 1.4 shows the resulting statistics. The comparison of pre- and postdebate, shows that the MSB for Beliefs decreases slightly (by 0.12 of a standard deviation), but not significantly so (*Mann-Whitney* test $H_0 : MSB_{Pre} = MSB_{Post}$, $p = 1.000$). When it comes to Attitudes, polarization actually increases slightly (by 0.05 of a standard deviation), but again without statistical significance.

To check whether our measure is capable of picking up the changes in polarization

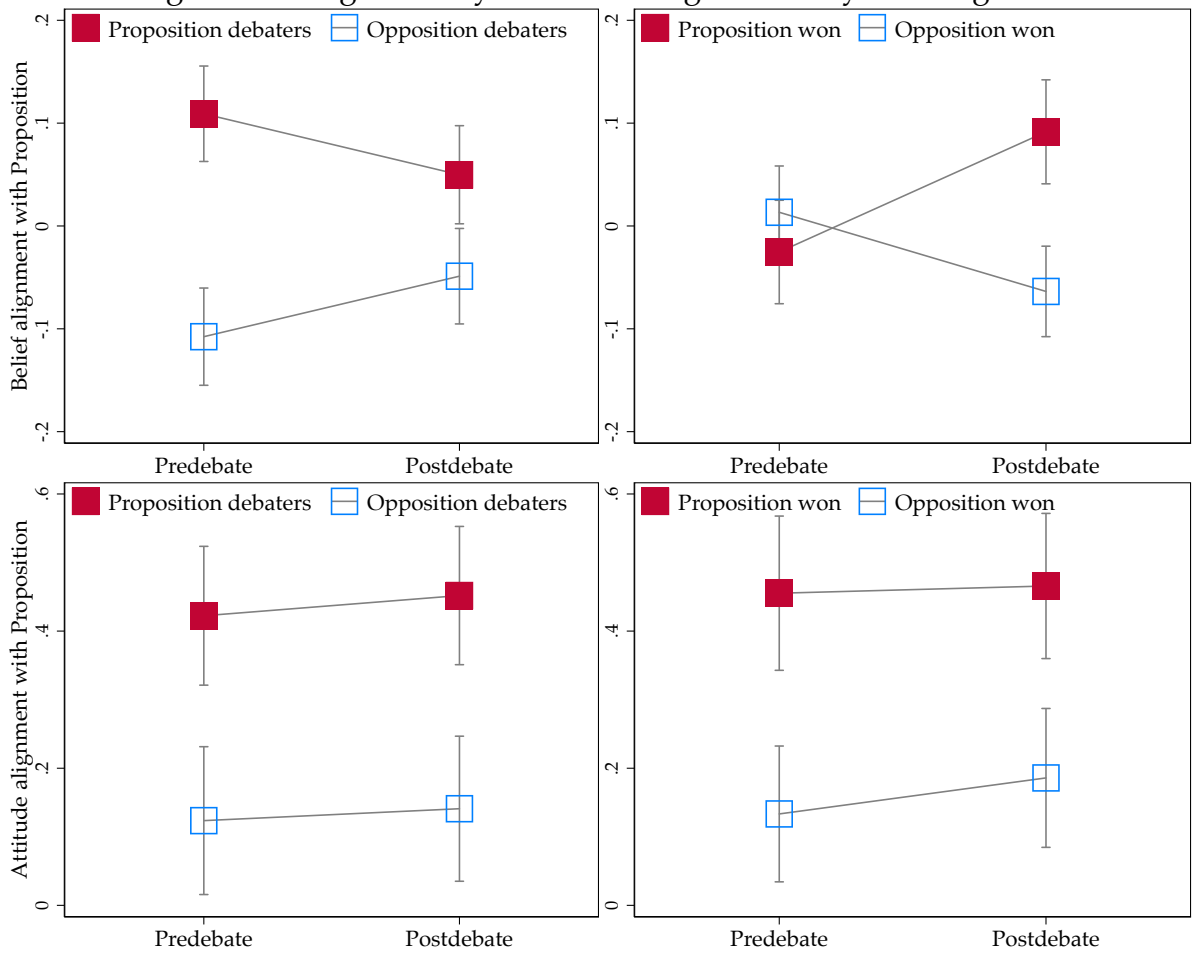
documented in the previous subsection (i.e. self-persuasion), we also include the polarization in Factual Belief at baseline. We see that the MSB for factual beliefs increases significantly from baseline to predebate (*Mann-Whitney* test $H_0 : MSB_{Base} = MSB_{Pre}$, $p = 0.023$), showing that the MSB measure captures the polarizing effects of self-persuasion. Moreover, it also increases between baseline and postdebate by 0.57 of a standard deviation (*Mann-Whitney* test $H_0 : MSB_{Base} = MSB_{Post}$, $p = 0.031$), showing that the overall debating experience leads to an increase in polarization.[14] To assess the robustness of these findings, Section A.3 applies two other prominent measurements of polarization in the literature based on Desmet, Ortuño-Ortín, and Wacziarg, 2017 and Duclos, Esteban, and Ray, 2004. These approaches lead to similar conclusions as our main analysis.[15]

Note that our variance decomposition does not take into account the *direction* of disagreement, so it ignores cases with reversed disagreement. To address this, we look at the dynamics of alignment with the Proposition for both treatment groups. The two left panels of Figure 1.5 show the data for Beliefs (top left panel) and Attitudes (bottom left panel). The debate roughly halves the gap in Beliefs from 21.7 percent ($Mann - Whitney$ test, $p = 0.001$) of a standard deviation to 9.9 percent ($Mann - Whitney$ test, $p = 0.112$). Thus, the effects of self-persuasion decline after the debate, but with a p-value of 0.112, we cannot accept the null hypothesis of equal beliefs either. With respect to the Attitudes, there is no reduction in disagreement, as alignment remains constant at about 0.3 donation ranks.

[14] Figure A.3 and Figure A.4 dissect the evolution of disagreement between debaters question by question, and demonstrate that polarization occurs on a broad range of issues.

[15] Mimicking the variance decomposition, (Desmet, Ortuño-Ortín, and Wacziarg, 2017)'s measure of cultural distance increases significantly from baseline to postdebate, and is reduced slightly from predebate to postdebate–although not significantly so. The polarization index by Duclos, Esteban, and Ray, 2004 shows that the polarization of factual beliefs appears stable through the three elicitations. This index however does not perform too well with survey responses that have a high mass of reports at focal points (e.g. for factual beliefs these are 0, 50, and 100). Distributions with (more than one) artificially strong modes are spuriously identified as substantially polarized, making relatively small changes in actual polarization hard to detect

Figure 1.5: Alignment by Position Assigned and by Winning Side



Note: Aggregating across all participants and rounds of debate, this chart plots averages for ours measures of belief and attitude alignment with the Proposition by different subgroups. These are constructed using the set of questions labeled C, D, E and F in Table 1.2 following the same procedure as for our regression analysis in Section 1.3.1 and Section 1.3.1. Left panels report averages for each of the two sides assigned in the debate. Right panels report averages across rooms where each of the two sides won the debate. Ranges indicate standard errors. Corresponding regression analysis in Table A.9.

Finally, we investigate whether the dynamics of polarization are related to emotions during the debate, as Mutz, 2007 shows that incivility during debates leads people to take opposing views less seriously. To measure emotions during the debate, enumerators recorded both a subjective measure of the "heatedness" of each debater, and the number of interruptions during the debate. The analysis in Section A.6 shows that debaters whose baseline beliefs are aligned with their persuasion goals also give more heated speeches, but greater heat in a debate does not moderate the convergence

of views (see Section A.3).

Persuasion during the Debate

We now turn to the impact of persuasion during the debate. To do so, we look at whether Beliefs and Attitudes move in the direction of the winning side, where the winning side is determined by the panel of judges based on the quality of arguments. The right panels of Figure 1.5 show the dynamics of Beliefs (top right panel) and Attitudes (bottom right panel) comparing rooms in which the Proposition won with rooms in which the Opposition won. Predebate Beliefs move towards the winning side postdebate. In particular, a Proposition win is accompanied by a 11.7 percentage of a standard deviation shift towards the Proposition. An Opposition win coincides with a 7.7 point shift toward the Opposition.

To have a clean measure of the amount of persuasion, we compare the belief movement between rooms with opposite winners using a difference in difference measure. Specifically, our persuasion measure tells us how much more aligned postdebate Beliefs are with the Proposition in a room where the Proposition won than in a room where the Opposition won, subtracting the differences in predebate Beliefs. On this measure, the persuasion effect in Beliefs is 19.4 percent of a standard deviation. Instead, for Attitudes, the corresponding effect is -0.04 donation ranks. That is, we see no persuasion effect at all, as Attitudes drift slightly towards the Proposition in both groups.

Based on these results, we can make a comparison between the effects of persuasion and self-persuasion. Average self-persuasion, as measured by the pre-debate gap between the Proposition and Opposition, is 21.7 percent of a standard deviation for Beliefs, and 0.30 donation ranks for Attitudes. In both cases, this is larger than the effect of persuasion defined and computed in the previous paragraph, and for Attitudes, only self-persuasion matters. Thus, when quantifying the outcomes of debates on Beliefs and Attitudes in our setting, self-persuasion effects dominate persuasion effects. These results are in line with Janis and King, 1954, who show that making people ar-

gue for a specific side is more effective in changing their views than hearing similar arguments from someone else. These findings can inform a literature in economics and political science about the circumstances under which persuasive communication can shape beliefs and preferences (DellaVigna and Gentzkow, 2010; Druckman and Lupia, 2016).

Result 4. *We find that debating has different effects on Beliefs and Attitudes.*

- *For Beliefs, debating reduces but does not eliminate the disagreement induced by self-persuasion. We see an effect of persuasion, as Beliefs shift towards the winning side.*
- *For Attitudes, debating does not have a discernible effect on disagreement, nor do we see persuasion effects.*

Overall, we find that the effect of self-persuasion is at least as strong as the effect of persuasion.

1.4 Mechanisms and Consequences of Self-Persuasion

We now discuss several secondary research questions. First, we delve deeper into the psychological mechanisms behind self-persuasion. We then discuss the relation between self-persuasion and debating success.

1.4.1 Psychological Mechanisms of Self-Persuasion

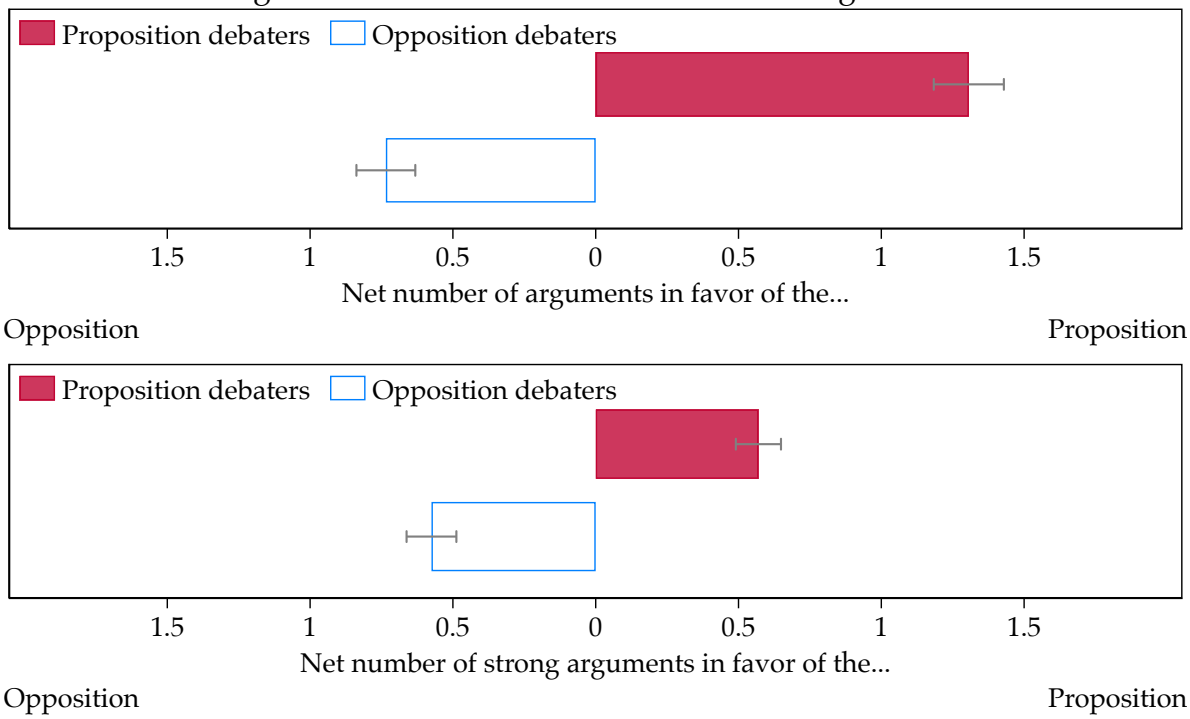
What psychological mechanisms underlie the self-persuasion documented in the previous section? Two plausible candidate mechanisms by which social interactions cause bias in beliefs and attitudes are self-deception and bounded rationality. Self-deception refers to a process of motivated reasoning in which debaters "choose" their beliefs. In this account, put forward in Von Hippel and Trivers, 2011, self-persuasion is a subconscious strategy aimed at increasing persuasiveness. It does so by reducing nervousness, give-away tells or other manifestations of doubt or cognitive dissonance arising from a discrepancy between one's persuasion goals and true beliefs. This theory has received support in recent laboratory studies (Smith, Trivers, and Hippel, 2017; Schwardmann and Weele, 2019; Solda et al., 2019).

By contrast, Mercier and Sperber, 2011 argues that self-persuasion results from bounded rationality or cognitive heuristics. In the process of preparing for a debate, debaters may naturally gather more arguments for their position than against it. If debaters then fail to take into account that arguments were generated in a biased fashion, then they may take the asymmetry of generated arguments as evidence for the strength of their position. Such "selection neglect" has been documented in multiple studies (Juslin, Winman, and Hansson, 2007; Barron, Huck, and Jehiel, 2019). Related ideas underpin the notions of "availability bias" (Tversky and Kahneman, 1973) and "persuasive argument theory" (Vinokur and Burstein, 1974), which maintain that the number, novelty or salience of arguments drive belief formation.

Selection neglect implies that if debaters generate more arguments on their own side of the debate, then this asymmetry will mediate self-persuasion. To test this, we asked debaters in the predebate survey for the number of arguments they came up with during their preparation time, both for and against the motion. We also asked them how many of these arguments they considered to be "very strong". Figure 1.6 shows the average net number of arguments debaters came up with on both sides by treatment. As is clear from the graph, debaters engage in asymmetric selection of arguments. On average, they come up with one additional argument and one half of a "strong" argument in favor of their own side.

To quantify the impact of this asymmetry, we conduct a parametric causal mediation analysis (Imai, Keele, and Yamamoto, 2010) - see Appendix A.8 for details. We define s_i, the number of aligned arguments as a fraction of total arguments considered during preparation time, and investigate how this mediates self-persuasion on our three main outcome variable. The results in Table 1.6 reveal that s_i drives between 29 percent and 57 percent of the self-persuasion effect. The fraction is largest for Confidence and smallest for Factual Beliefs.

Figure 1.6: Differences in the Number of Arguments



Note: Ranges indicate standard errors.

These results suggest that selection neglect plays an important role in self-persuasion, but that mechanisms of self-deception are about equally, if not more, important. The quantitative result is subject to some uncertainty: on the one hand, we cannot rule out that selection neglect is itself (partially) driven by self-serving motives (Exley and Kessler, 2019), leading to a possible overestimation of the importance of the heuristic explanation. On the other hand, our measures of the number of arguments may be underestimated due to measurement error.[16] While more quantitative evidence is therefore needed, the results support the idea that both mechanism have a role in self-persuasion.

[16]See also Section A.8, where we discuss (i) the *sequential ignorability* assumption needed to identify causal mediation effects, and (ii) measurement error potentially attenuating the estimates of these effects (Cessie et al., 2012).

Table 1.6: Decomposition of Treatment Effect in Mediated and Direct Effect

	Beliefs	Attitudes	Confidence
Average causal effect mediated by s_i (ACME)	0.058	0.158	2.340
	(0.045)	(0.075)	(1.131)
Average direct effect (ADE)	0.143	0.129	1.714
	(0.075)	(0.156)	(1.854)
Average treatment effect (ATE)	0.201	0.287	4.110
	(0.066)	(0.137)	(1.558)
ACME/ATE	0.289	0.551	0.569

Note: Estimates obtained following the procedure outlined in Appendix D of Imai, Keele, and Tingley, 2010: we estimate the Linear Structural Equation Model using random effects regressions with the full set of controls as in Section 1.3.1, and we use the estimated sampling distributions to draw 100 simulations of potential mediators and potential outcomes. We average the differences of potential outcomes across the 100 simulations to obtain an estimate of the mediated effect. We repeat the procedure 1000 times from bootstrap samples to obtain standard errors of the estimates.

Finally, we can rule out several other mechanisms for self-persuasion that have been proposed in the literature. First, the randomization of persuasion goals excludes the priming of political affiliations (e.g. Petersen et al., 2013) or confirmation bias (e.g. Fryer, Harms, and Jackson, 2018). Second, Falk and Zimmermann, 2016 propose that the consistency of opinions and arguments may be a signal of intellectual skill. In our setting, the anonymity of the surveys rules out that subject engage in such signaling. Third, given the high levels of intrinsic motivation and short-timeframe, it is unlikely that people self-deceive to overcome time-inconsistent preferences (Bénabou and Tirole, 2002). Fourth, subjects had very little opportunity to acquire new information, and thus engage in selective search from external sources (Taber and Lodge, 2006). Finally, debaters are unlikely to actively think about research hypotheses and bias their responses accordingly, since the randomization is such a natural part of the tournament. Section A.7 provides more analyses that rule out experimenter demand effects.

1.4.2 Self-Persuasion and Debating Success

We now turn to the relation between self-persuasion and success in the debating competition. This relation is of interest for two reasons. First, it can inform our view of the psychological mechanisms underlying self-persuasion that we discussed above. A negative relation with debating success is consistent with an explanation of self-persuasion in terms of cognitive errors. By contrast, a positive relation is in line with strategic self-deception, where cognition is optimized for persuasiveness. Second, the success of self-persuasion in the context of a debating competition may tell us something about its prevalence in broader contexts. If self-persuasion is detrimental to persuasiveness, it would be less likely to constitute a widely observed phenomenon. However, if self-persuasion is not detrimental to persuasiveness, we might expect it to be common, even for people, such as politicians, whose professional success relies on persuasion.

Unfortunately, our dataset is not ideally suited to look at the causal effect of self-persuasion. The ideal experiment would create exogenous variation in self-persuasion. However, this would require changing debating objectives and procedures, which was not possible at such high profile competitions. Nevertheless, correlations may give us a valuable input for future research. Moreover, we can exploit the alignment of factual beliefs at baseline, which is random, to look at the effect of belief alignment on persuasiveness.

Is self-persuasion more prevalent among successful debaters? If successful debaters are more likely to engage in self-persuasion, we should expect a positive interaction effect between debater success and self-persuasion. To look at this, we add an interaction term to the regression model 1.1, used to study self-persuasion on all our three outcomes. Debater success is measured by "achievements" – the number of semi-finals reached by debaters in international tournaments–elicited in the baseline survey before treatment. Table 1.7 presents the results of such estimation. In each regression, we control for debating experience by including the number of years a debater has been

active.

Table 1.7: Panel Regressions for Heterogeneous Effects of Persuasion Goals

	Factual Beliefs		Attitudes		Confidence	
	(1)	(2)	(3)	(4)	(5)	(6)
Debater in proposition	0.203***	0.229***	0.300**	0.211	4.319***	2.784*
	(0.062)	(0.070)	(0.145)	(0.167)	(1.554)	(1.640)
Debater in proposition × Achievements		-0.007		0.024		0.419*
		(0.011)		(0.033)		(0.255)
Socio-demographic and experience controls	✓	✓	✓	✓	✓	✓
Round fixed effects	✓	✓	✓	✓	✓	✓
Observations	851	851	850	850	850	850

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level. Socio-demographic controls include age, gender, and an indicator for whether the debater's nationality is from the country that hosts the competition. Experience controls include the number of years the debater has been actively debating.

The results in column 1 indicate that self-persuasion on factual beliefs is not related to success in past tournaments: more and less successful debaters engage in self-persuasion to a similar extent. Though not (highly) significant, we find higher estimates for the interaction term for attitudes (column 4, $p = 0.471$) and confidence (column 6, $p = 0.100$). For debaters who have never made it to the semi-finals of an international tournament we estimate that for these variables the self-persuasion effect is 30 and 35 percent smaller, respectively.

Does belief and attitude alignment help persuasiveness? We analyze whether judges' evaluations of debaters' persuasiveness correlate with the alignment of debaters with their persuasion goal. We have four measures of a debater's alignment with the persuasion goal: Factual Belief alignment at baseline, Factual Belief alignment at predebate, Attitude alignment at predebate, and Confidence in Proposition at predebate. Note that only the first of these measures counts as exogenous variation, as it was measured before the treatment was administered. As measures of persuasiveness in

the tournament we have both a broad persuasiveness score provided by each judge independently, as well as a technical score of the quality of debater's arguments that is given by judges in agreement after the debate is over.

Table 1.8: Pearson's Correlation Between Persuasion Outcomes and Alignment Variables

	Broad persuasiveness	Quality of arguments
	(1)	(2)
Baseline belief alignment	-0.006	0.035
	(0.859)	(0.302)
Predebate belief alignment	-0.019	0.025
	(0.572)	(0.451)
Predebate attitude alignment	0.181	0.041
	(0.590)	(0.228)
Predebate confidence in own position	0.006	0.019
	(0.851)	(0.571)
Observations	883	883

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: P-value for statistical significance in parentheses. Broad persuasiveness is evaluated by each judge on a panel independently; but we average the individual scores at the debater-round level. Alignment variables transform our main raw outcomes as in Section 1.3.1, and change the sign of these outcomes for opposition debaters to obtain variables that become larger (smaller) as the debater exhibits greater (less) alignment with their persuasion goal.

Table 1.8 presents correlations between our measures of alignment and persuasiveness across all rounds of debate. None of our alignment measures is a significant predictor of persuasiveness. One explanation for this null result is that measurement error attenuates the relations between the variables. In fact, while alignment with the persuasion goal may be partially or wholly captured using Factual Beliefs, Attitudes, and Confidence, actual debater's alignment remains a latent variable. In addition, the low inter-rater agreement between judges (*Cohen's Kappa* $= 0.083$) on the broad persuasiveness of each debater also raises concerns regarding the overall quality of judges'

unincentivized responses.[17]

In summary, although we find slightly more positive than negative point estimates, there are only weak correlations between debater success and the alignment of their attitudes and beliefs with their persuasion goal. The available variation in our dataset does not allow definite conclusions about the relation between self-persuasion and debater success. This remains an important area for future research.

1.5 Conclusion

Our data show that people distort their factual beliefs, attitudes and confidence in the direction of the position they are randomly assigned to argue. Debaters engage in such self-persuasion and start believing in "alternative facts" despite incentives for accuracy and exposure to opposing views. These results obtain in prestigious tournaments in a sample that is a regular supplier of future elites and politicians. We find no evidence that self-persuasion is detrimental to success and no reason to suspect that it disappears with experience, so our findings are likely to apply to professionals in a number of fields and applications.[18]

Here we enumerate a number of applications where our results matter and provide impetus for future research. First, they speak to the institutional foundations of deliberative democracy. Self-persuasion and polarization obtain in a competitive setting that mimicks the British parliamentiary institution, calling into question the power of debate to bridge conflicts in society. These results need not obtain in more cooperative settings, where an agreement needs to be reached among parties. Recent research shows that prompting people to focus on the opposing side of the argument or to arrive at shared solutions can lead to more balanced argumentation (Felton, Crowell, and Liu, 2015; Perkins, 2019). It is therefore an important question how the debating context affects self-persuasion, and how it can be designed to promote convergence of

[17]The *Cohen's Kappa* coefficient ranges between 0 (expected level of agreement that can be obtained by chance) and 1 (perfect agreement).

[18]Our results do not imply that debaters are especially prone to self-deception or factual misperceptions. We encountered engaged and impressively knowledgeable individuals at the debating competitions. The extent to which these features make individuals more or less prone to self-persuasion remains an open question for future research.

views and a shared understanding of facts.

Second, self-persuasion can explain instances of polarization in political contexts where convincing others is of central importance. For instance, self-persuasion offers a reason why polarization is more severe in the US congress than it is in the American public (Fiorina and Abrams, 2008), why it is so strong on social media platforms, especially if people are exposed to opposing views (Bail et al., 2018), why greater engagement with the political process causes greater and persistent polarization (Mullainathan and Washington, 2009), and why people who joined the Republican party exclusively for their view on abortion then saw their other beliefs fall in line with the party (Gould and Klor, 2019). It also suggests alternative or additional motives for political behavior than are commonly assumed. For instance, canvassing and proselytizing activity may be important not just to grow the base, but also for deepening the convictions of existing followers Gal and Rucker, 2010. Similarly, opportunistic political U-turns or flip-flops may be the cause of genuine conversion in the process of defending the new position.

Third, self-persuasion offers insights for markets with asymmetric information. It predicts that sellers in economic transactions risk "drinking the kool-aid" and become overly optimistic about their product. This may explain why financial advisors privately invest in the under-performing funds for which they receive sales commissions (Linnainmaa, Melzer, and Previtero, 2018). It may also be a driving force behind the development of asset market bubbles, for instance during the financial crisis of 2007-8, where private real-estate portfolios of agents working in sales departments of mortgage providers under-performed those of other agents as well as non-specialists (Cheng, Raina, and Xiong, 2015). Self-persuasion also offers insight into the sometimes spectacular rise and fall of start-up companies like Theranos, as it predicts that entrepreneurs trying to lure investors are likely to become overconfident and miscalibrated.

More generally, and perhaps most importantly, we show that social interactions invite systematic deviations from the Bayesian ideal, still a mainstay of economic the-

ory. Our findings lend support to theories that reserve a fundamental role for social influence and persuasion in the development and operation of our cognitive capacities (Von Hippel and Trivers, 2011; Mercier and Sperber, 2011), and can provide a unified explanation of non-Bayesian cognitions that are currently being studied separately in the field of behavioral economics. They help explain why people engage in various self-enhancement strategies and become overconfident about their abilities (Trivers, 2011; Schwardmann and Weele, 2019), why they are more eager to confirm than to disconfirm their views (Nickerson, 1998; Benjamin, 2019), why they look for exculpatory narratives and exploit wiggle room in moral dilemmas (Dana, Weber, and Kuang, 2007; Exley, 2015; Di Tella et al., 2015), and why they appear conveniently unaware of their darker motives (Kurzban, 2012; Simler and Hanson, 2017).

Further research is necessary to test the explanatory power of self-persuasion and the interactionist approach in these domains. Our findings raise expectations that such a research program will lead to substantial revisions in the standard view of human cognition, a view eloquently expressed by John Maynard Keynes. When accused of inconsistency, he purportedly responded: "When the facts change, I change my mind. What do you do Sir?". For many people the answer appears to be "the reverse".

Chapter 2

Social Influence in Prosocial Behavior: Evidence from a Large-Scale Experiment

2.1 Introduction

The increasing social connectivity of modern times fosters opportunities for social interactions and comparisons with others. A growing literature illustrates how information and cues about the behavior of others can induce social influence: the effect of others' actions on individual behavior. Social influence plays an important role across a broad range of domains that includes charitable giving (Frey and Meier, 2004), financial decision making (Bursztyn et al., 2014), marketing (Bapna and Umyarov, 2015), political participation (Cantoni et al., 2017), tax evasion (Drago, Mengel, and Traxler, 2020), and well-being (Aral and Nicolaides, 2017). While in most social influence studies individuals *observe* others' behavior, various models (e.g. Bernheim, 1994; Akerlof, 1997) explain the spread of social influence even for unobservable behavior via *conformity*. Isolating this behavioral mechanism requires ruling out the learning opportunities derived from observing the actions of others.

In this paper, we study social influence in prosocial behavior through conformity, i.e. when actions are *not* directly observable. Social influence makes actions of connected agents interdependent, but such interdependencies are often ignored in standard models of prosocial behavior.[1] We examine how information about others' environments generates social influence. We intentionally shut out observability. This allows us to disentangle the mechanisms of social influence and to assess the scope of social influence in applications where information about the behavior of others is harder to access compared to information about the constraints, incentives or institutions that others face.

We analyze social influence through a conceptual framework and an experimen-

[1] Much of the theoretical literature models prosocial behavior and public good contributions as games of strategic substitutes. The most prominent examples of such theories are represented by models of pure altruism (Becker, 1974) and impure altruism (Andreoni, 1989; Andreoni, 1990).

tal design focused on a notion of conformity that is due to identification with a peer and her motives. When individuals encounter someone doing good out of intrinsic motives, they are inspired (or compelled) to conform to this behavior, and deviating creates a psychological cost. Given this preference, agents use the economic environment to infer intentions of the social reference and attempt to conform to their behavior even when this is not observable.

In a large-scale online experiment, 2,914 individuals engage in pairwise interactions before they independently take part in a real effort donation task. The two main outcomes of interest are (i) the amount of charitable donations individually generated through the donation task and (ii) expectations of the amount generated by the other player within the pair. In our experiment, individuals can generate donations to a charity through a tedious physical task. We experimentally manipulate private incentives for making donation: for each player in a pair, we cross-randomize one of three levels of piece-rate (*zero*, *moderate*, and *high*) private incentives to generate donations for *Médecins Sans Frontières*. Variation in the incentives of the other player in the pair allows us to uncover social influence among peers: if an agent cares to conform, an increase in her peer's incentives will have both a direct effect on her peer's donations and an indirect effect on the agent's donations as she tries to minimize distance with the actions of her peer. We can then identify the social influence effects of peer incentives and evaluate different behavioral motives by estimating the contemporaneous effect of peer incentives on both expectations—about donations of the peer—and donations of the agent whose incentives are held constant. Before the treatment manipulation, pairs of subjects participate in a joint problem solving task, which we adopt to induce social proximity between paired players (Chen and Li, 2009; Chen and Chen, 2011) and increase relevance of the peer as a social reference. After that, we elicit a survey measure of social proximity (Cialdini et al., 1997), which we use to investigate how social proximity determines propagation of social influence.

We find evidence of social influence in donations: when the peer's incentives increase from *zero* to *moderate*, subjects expect their peer to increase donations and in

turn, they donate more themselves. These effects are entirely driven by subjects who exhibit a close social connection to their peer, for whom the effect of increasing the peer's incentives from *zero* to *moderate* on donations is as large as half the effect of increasing *their own* private incentives from *zero* to *moderate*. However, when the peer's incentives further increase from *moderate* to *high*, we find a different result: individuals correctly expect their peer's donations to not be affected by higher incentives, and they themselves donate less when their peer has *high* incentives. Thus, individual donations respond non-monotonically to peer's incentives. These effects are, again, entirely driven by the subsample of individuals who feel socially close to their peer. For individuals who do not feel close to their peer, we cannot reject the null of no social influence.

We propose a mechanism related to Fuster and Meier (2009), and argue that the strength of the desire to conform depends on whether the peer engages in the behavior for non-selfish reasons. Higher incentives for the peer can thus have an ambiguous effect on behavior. If incentives are "too generous", the peer's behavior may no longer be viewed as non-selfish, and the desire to conform weakens. Thus, individuals may well reduce effort in response to higher incentives for their peer. We formalize this intuition in a model and show that non-monotonicities as observed in our experiment can be generated in a simple version of the model.

One might suspect that higher peer incentives reduce an individual's contributions due to substitution effects from models of impure altruism. However, for such substitution to occur, it needs to be the case where the individual expects a higher contribution from her peer. In our setting, this is clearly not the case.

Differences in incentives between individuals may also give rise to incentive-inequality effects described in Breza, Kaur, and Shamdasani (2017). Incentive inequality in that sense predicts that, conditional on own incentives, donations decrease with the difference in incentives to the peer. Thus, they predict a monotonicity with regard to that gap. However, our non-monotonicities arise for all levels of own incentives. In most of these cases, incentive inequality would predict the opposite pattern. We develop a

formal test and can reject that incentive inequality explains the pattern we find.

Our work broadly contributes to the large literature in economics and psychology that has studied empirically whether social information can produce social influence on prosocial behavior, both in the lab (Cason and Mui, 1998; Bohnet and Zeckhauser, 2004; Eckel and Wilson, 2007; Krupka and Weber, 2009; Servátka, 2009; Duffy and Kornienko, 2010; Bigenho and Martinez, 2019) and in the field (Frey and Meier, 2004; Shang and Croson, 2009; Chen et al., 2010; Fellner, Sausgruber, and Traxler, 2013; Cantoni et al., 2017; Bruhin et al., 2020). Our main contribution to this literature is to show that observing the behavior of others is not necessary for people to be subject to social influence. In fact, they will attempt to infer how others behave and conform to that behavior.

We also contribute to a growing literature that tries to disentangle the mechanisms of social influence. While we are not the first that try to separately identify social learning from conformity (Bursztyn et al., 2014; Lahno and Serra-Garcia, 2015; Gilchrist and Sands, 2016), our experiment is, to the best of our knowledge, the first with a focus on conformity in an environment that entirely removes any opportunity for social learning. Moreover, compared to these papers, we are the first to study conformity in the prosocial domain: Bursztyn et al., 2014 investigate social learning and the shared experience of holding an asset as distinct mechanisms of peer effects in financial decisions; Lahno and Serra-Garcia, 2015 isolate conformity in lottery choice through a decision environment stripped down of complexity to minimize the scope for social learning; Gilchrist and Sands, 2016 use weather instruments to estimate the effect of cumulative movie viewership on the probability of going to watch a movie and run various robustness checks to rule out social learning about quality of the movie.

Our findings have implications for a large literature on social influence and incentives for charitable giving and volunteering (e.g. Eckel and Grossman, 2003; Landry et al., 2006; Huck, Rasul, and Shephard, 2015; Meer, 2017; Perez-Truglia and Cruces, 2017), furthering our understanding of the forces that modulate the channels of social influence. It enriches the literature on the damaging role of incentives on norm-

adherence (Gneezy and Rustichini, 2000a; Gneezy and Rustichini, 2000b; Fuster and Meier, 2009), by demonstrating a more nuanced role of incentives. Furthermore, we add, to an empirical literature on the role of social proximity in social influence mediated by social information see e.g. Topa, 2001; Leider et al., 2009; Bond et al., 2012; Dimant, 2018, evidence that social proximity also modulates social influence in the absence of social information. This evidence is important because it shows that social proximity matters even when benefits of (and opportunities to punish in) future interactions are absent.

Most closely related to ours is the work of Kessler, 2017, who provides field and laboratory evidence that public endorsement of peers to a charitable cause can produce large complementarities in giving even when the actual amount of money donated is not observable. He proposes social learning and conformity as primary behavioral channels to explain such findings. Our work complements this paper in two important ways: First, Kessler, 2017 shows that endorsements affect beliefs about the quality of a charity and others' donations. Our experiment is designed to hold constant beliefs about the quality of the charity to make a first attempt at separately identifying conformity from social learning in the prosocial domain. Second, we use a novel approach to identify social influence based on private incentives to donate in newly-formed social bonds. This allows us to learn new lessons about the interaction between prosocial motivations, social proximity and conformity.

The remainder of this paper is organised as follows. Section 2.2 presents the experimental design and predictions. Section 2.3 illustrates the results and discusses mechanisms of social influence. Section 2.4 concludes.

2.2 The Experimental Setup

2.2.1 Experimental Design

We conduct an online experiment with registered workers from Amazon Mechanical Turk. The study develops over five stages, featuring a full 3×3 between-subject design plus an additional control treatment. All subjects in the experiment are randomly

grouped into pairs. Prior to learning about the main experimental task, subjects make contact with the other player in the pair. Pairs are formed after Registration, and the first three stages are common to all pairs. In the fourth stage, each pair is randomly assigned to one of ten treatments. The experiment concludes with a short survey and review of the payoffs. We present each stage in detail below.[2]

1. Registration. Invited subjects accept the general conditions for participating in the experiment before accessing the software interface. The study begins with general instructions that outline the key stages of the experiment: subjects are informed that they will be randomly paired with another player with whom they will jointly complete the first task, followed by the second task to be completed independently. After reading the initial set of instructions, each subject chooses a number from 1 to 6, which they are told will matter for the variable component of their pay at the end of the experiment. We introduce *tokens* as the experimental currency. This stage is concluded by a short survey to collect demographic information (i.e. name, gender, age, and experience on Amazon Mechanical Turk), which subjects are told will only be shared with their peer.[3]

2. Joint problem solving task. As subjects progress to this stage of the experiment, pairs are formed at random and subjects are introduced to their peer: they are presented with the demographic information of their peer (i.e. stated name, gender, age, country of residence, and experience on Amazon Mechanical Turk) on their computer screen. [4] All our subjects are residents in the United States.

Similar to Chen and Li, 2009, we use a joint problem solving task to favor the formation of a social connection between paired players. In this task, pairs of players see the same four famous paintings. For each painting, subjects are incentivized to identify – in coordination with their peer – the corresponding artist from a list of five: each subject in the pair earns 20 tokens each time *both* players give the correct artist for the

[2]Full experimental instructions can be found in the supplemental material, Appendix C.

[3]We cannot verify that this information is truthfully provided. We ask people to provide a name to facilitate interactions, but we did not expect players to recognize the peer as acquaintance/friend. Chat scripts provide no evidence of pre-existing relationships among paired participants.

[4]The order of arrival to this page constitutes our random matching protocol.

same painting.[5] Paired players can solve the task through a private online chat (see interface in Figure B.1). We differ from Chen and Li, 2009 by making rewards dependent on both own and peer's answers to increase incentives for establishing social contact. Payoffs are revealed at the end of the experiment.

3. Oneness elicitation. We measure social proximity with the *oneness* scale. There are two main reasons why this is a natural choice for the study: The *oneness* scale has been found to explain social proximity for dyadic relationships relatively well in comparison to more involved questionnaire-based scales from social psychology (Gächter, Starmer, and Tufano, 2015), and it is fast and simple to administer (see Figure B.2). The oneness scale was first proposed by Cialdini et al., 1997 as a simple mean of two underlying scores: (i) the Inclusion of Other in the Self (IOS) scale and the (ii) WE scale. The IOS scale (Aron, Aron, and Smollan, 1992) is an easy-to-administer pictorial measure of social proximity between the research subject and a related person, constructed by simply asking subjects to indicate which of seven diagrams, composed of two increasingly overlapping circles, best represents their connection to the related person of interest. Cialdini et al., 1997 later proposed to integrate the IOS scale with the WE scale, which asks subjects to express the extent to which they would refer to themselves and another person of interest as *we*, to capture complementary aspects of group membership embedded in social relationships. Both scales are elicited without incentives.

4. Donation task. For this task, subjects have to decide how many donations to generate for charity and make a point prediction about the number of donations their peer will generate. We treat such point prediction as proxy of beliefs of peer's giving.[6] To limit the scope of anchoring effects, we elicit expectations and desired number of donations simultaneously. After recording the two variables, subjects carry out the real effort task that generates these donations. Each donation requires entering 100

[5] We make the task hard by listing possible artists from relatively similar epoch and style.

[6] For practical reasons we do not elicit the entire belief distribution, but instead use a measure that most likely captures the perceived mode of giving of the peer. To limit the scope for motivated reasoning, we incentivize correct predictions with a 20 tokens prize.

sequences of keystroke combinations "w"-"e" on a computer keyboard.[7]

Prior to eliciting beliefs and donations, subjects go through a small training exercise to familiarize themselves with the real effort task, and this allows us to screen out subjects who are unable to solve the task. Thereafter, the software randomly assigns pairs of subjects to one of the ten different treatments.

Our experimental treatment manipulations simultaneously vary incentives to behave prosocially for both subjects in a pair. To make it very clear that variation in monetary incentives is random and independent between peers, all players in the nine incentivized treatment conditions are provided with ex-ante identical lottery incentives. This is also important for ensuring that different incentives could not be viewed as a signal for the importance of the task (Ellingsen and Johannesson, 2008). Subjects earn 50 tokens for each donation generated if the number picked in *stage 1* matches the roll of a fair die. Across incentivized treatments, we vary the *expected* stakes of monetary incentives for each player by means of a simple information device. The device randomly determines whether to disclose if the matching die has a face number between the largest three or the smallest three figures of a die. When this signal is provided, depending on the initial number chosen, this either reduces to zero the chances of getting the piece-rate incentive to generate donations (incentives are *zero*), or it increase chances to 1 in 3 (incentives are *high*). When this signal is not provided, the probability of getting the piece-rate incentive for generating donations is not updated and remains 1 in 6 (incentives are *moderate*).[8] To make incentives common knowledge within each pair, we reveal to subjects their peer's signal and initial chosen number. We also make sure that subjects understand both their own and peer's incentives by (i) framing as "lucky" ("unlucky") the die roll when incentives are *high*

[7]We choose a sterile task to limit the scope for confounding factors. A similar task has been used in other experiments studying incentives for charitable giving (Ariely, Bracha, and Meier, 2009; Meyer and Tripodi, 2017), and effort provision (DellaVigna and Pope, 2016; DellaVigna and Pope, 2017).

[8]We prefer this probabilistic approach over randomizing a deterministic piece rate to reduce disappointment in pairs where one subject receives no incentive and her peer receives high incentives. The main disadvantage is that it potentially introduces subjective evaluations of probabilities (see e.g. page 637 of DellaVigna, 2018, for a discussion of the mixed evidence on probability weighting in real effort experiments). However, this approach has the great advantage that, by reducing disappointment, it helps avoid differential attrition across treatments.

(*zero*) and (ii) directly providing them with the updated probabilities of receiving the piece-rate to generate donations (see Figure B.3 for an example). This information revelation scheme produces variation in the magnitude of expected incentives for acting prosocially, for both player i and peer j of each pair, in a full 3×3 between-subject design. We enrich this design with a control *no lottery* condition. Figure 2.1 schematizes the experimental design.

5. Exit. In the final stage, subjects answer some unincentivized questions to check comprehension. The summary of individual payoffs concludes the experiment.

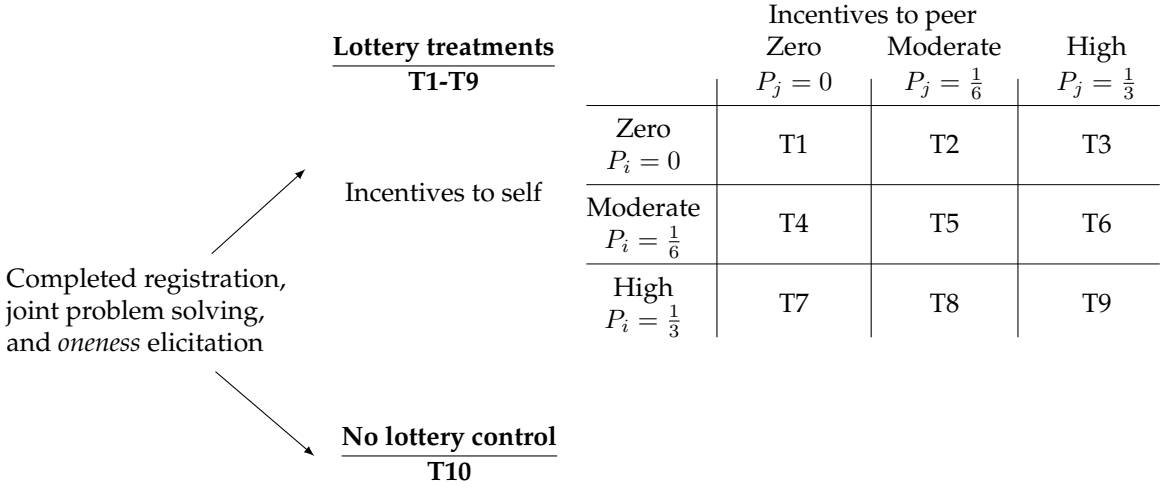


Figure 2.1: Overview of Experimental Design and Treatment Assignment

2.2.2 Conceptual Framework and Predictions

To formalize our strategy for identifying social influence, consider the following simple model of prosocial behavior. Two agents $a = \{i, j\}$ are presented with the opportunity to choose a donation effort d_a. There are four components to their utility: donations create at a monotonically increasing and convex private cost $c(d_a)$. Personal benefit from donations is a heterogeneous altruism component v_a per unit of d, distributed according to c.d.f $F(v_a)$ in the population, and a monetary benefit m. Agents have a preference (or feel pressured) to conform to their peer. We follow Sliwka, 2007 in assuming that people conform to the *natural* behavior of their peer d_j^n, which is j's

behavior absent pressures to conform.[9] That preference is captured by a loss function $\kappa_{i,j}(\cdot)$ that is convex, monotonically increasing in the absolute distance between d_i and the expected d_j^n (because there is heterogeneity in v_j). [10] We write the utility of agent i from contributing d_i as:

$$U(d_i|m_i, m_j) = (v_i + m_i)d_i - c(d_i) - \kappa_{i,j}(|d_i - E(d_j^n|\mathcal{A}_j(m_j)|) \tag{2.1}$$

where $E(d_j^n|\mathcal{A}_j(m_j)) = E_{v_j}\big(\underset{d_j}{\text{argmax}}\ (v_j+m_j)d_j - c(d_j)|v_j \in \mathcal{A}_j(m_j)\big)$. We use this model to understand contributions to large charities, for which changes of a few dollars in aggregate donations are the proverbial drop in the ocean. Hence we consider a model in which the marginal altruistic utility from donating to the charity is constant, but the model can certainly be extended to allow for decreasing marginal returns.[11]

The key feature of our model is the function $\kappa_{i,j}(\cdot)$. It combines the standard forces of conformism (Akerlof, 1997; Bernheim, 1994) with the innovation that the strength of conformity depends on how normatively "attractive" the role played by the peer is (Kelman, 1961). A role is normatively attractive if an agent desires to identify with it. We model this by assuming that an individual's cost from deviating depends on whether her peer engaged in the behavior for non-selfish reasons. We specify this as

$$\kappa_{i,j}(|d_i - E(d_j^n|\mathcal{A}_j(m_j)|) = -\frac{\lambda_{i,j}}{2} Pr(\mathcal{A}_j(m_j))(d_i - E(d_j^n|\mathcal{A}_j(m_j)))^2$$

with $\mathcal{A}_j(m_j) = \{v_j \in V : c(d_j^n)/d_j^n > m_j\}$. Thus, the set $\mathcal{A}_j(m_j)$ represents all agents for whom choosing d_a given the monetary incentive m_j does not cover their cost of effort. The more non-selfish types there are, the stronger the conformism the individual feels towards that behavior. The parameter $\lambda_{i,j} \geq 0$ measures the importance of conformity costs relative to the marginal utility of money, and may vary between individuals depending on how socially close they feel to each other (Bond et al., 2012;

[9] This formulation shuts out second-order strategic effects. It considerably simplifies the analysis, as it turns the solution into a maximization problem.

[10] We also normalize $\kappa_{i,j}(0) = 0$.

[11] In the appendix, we consider a model of impure altruism with diminishing marginal utility. We show that it predicts that an agent's donation are globally declining in her peer's incentives.

Gioia, 2017).

For the case of quadratic costs $c(d)$, it is easy to show that $\Pr(\mathcal{A}_j) = 1 - F(m_j)$, i.e. the fraction of non-selfish types is the density to the right of m_j in the distribution of altruism parameters $F(v_a)$.[12] In this case, the objective function simplifies to

$$U(d_i|m_i, m_j) = (v_i + m_i)d_i - \frac{cd_i^2}{2} - \frac{\lambda_{i,j}}{2}(1 - F(m_j))(d_i - E(d_j^n|m_j))^2 \tag{2.2}$$

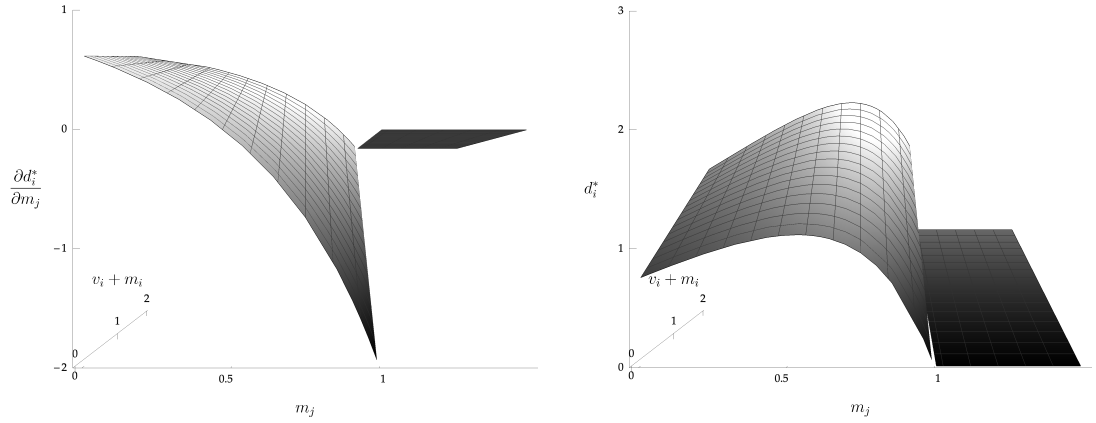
This yields the first-order condition that implicitly defines the optimal d_i

$$d_i = \frac{v_i + m_i + \lambda_{i,j}(1 - F(m_j))(d_i - E(d_j^n|m_j))}{c} \tag{2.3}$$

The equation illustrates how changes in m_j act through two channels on the individual's optimal behavior. The "traditional" conformism effect (Akerlof, 1997; Bernheim, 1994) acts through $E(d_j^n|m_j)$: higher incentives to j increase the normal effort d_j^n and thus act to increase d_j in equation (2.3). The second channel acts through composition effects: higher m_j reduces the fraction of individuals $1 - F(m_j)$ who engage in the behavior for non-selfish reasons in equilibrium. Thus, while the traditional conformism channel is unambiguously positive, the second channel acts against this and can overturn the sign of the overall effect.

In Figure 2.2, we illustrate the predictions of this model when $v \sim \mathcal{U}[0, 1]$ at varying levels of the private benefits to contribute $v_i + m_i$. The left panel shows how j's incentives m_j have positive effects on i's donations when incentives are low; these effects are decreasing in j's incentives and tend to become negative when m_j becomes large relative to $v_i + m_i$. When incentives are sufficiently large that the set of agents who engage in the prosocial activity out of altruism is empty, agents feel no need to conform and changes in m_j have no effect on d_i. These patterns translate into the non-monotonic relationship between m_j and d_i that is illustrated in the right panel of the figure.

[12] Because quadratic costs imply $d_j^n = \frac{v_j + m_j}{c}$, it follows that $\Pr(\mathcal{A}_j) \coloneqq \Pr(c(d_j^n)/d_j^n > m_j) = \Pr\left(\frac{v_j + m_j}{2} > m_j\right)$.



Note: The left panel graphs the marginal effects of increasing the peer's incentives (m_j) on the agent's donations (d_i) at different levels of the agent's private benefit to donate ($v_i + m_i$). The right panel graphs the agent's donations (d_i) as a function of her own private benefits to donate ($v_i + m_i$) and the peer's incentives (m_j).

Figure 2.2: Own donations as a function of own and peer's monetary incentives

In Section B.1.1, we study the model in a more general setting and show that for a general distribution of types $F(v)$ and a large set of cost functions, with constant elasticity of effort $k \leq 1$, there exists a threshold for $\tilde{m}_j$ above which i becomes unresponsive to changes in the incentives of her peer.

This theoretical framework offers two approaches to identify conformity through incentives. The first, less data demanding, hinges on estimating the indirect effect of changes in j's incentives to donate on i's donation behavior: conformity predicts that an increase in j's incentives should increase i's donations. The second, identifies the strategic complementarities of conformity by considering the effect of changes in j's incentives on both i's expectations about j's donations and i's donations: if donations are affected by conformity, changes in j's incentives shift both i's beliefs about j's donations and i's donations in the same direction.

The framework also provides an explanation for why not all actions of a social reference may lead to conformity in the same way. Much like in theories of prosocial behavior with incentives, e.g. for social signaling (Benabou and Tirole, 2006) and peer punishment (Dutta, Levine, and Modica, 2018), the extent to which agents wish

to adhere to the behavior of a social reference can be endogenous to incentives. Our experimental design allows us to separate conformity from these alternative explanations, to be discussed in section 2.3.4.

2.2.3 Procedures

To uncover the role and determinants of the conformity channel of social influence, we conduct six sessions of the experiment in 2017, between July 30 and August 4, recruiting 3,467 subjects on Amazon Mechanical Turk.[13] This is an online platform that is becoming increasingly popular for conducting economic experiments (DellaVigna and Pope, 2016) where thousands of registered workers are commonly employed in tasks that require human intelligence. Compared to lab subjects, workers on this platform are more heterogeneous in terms of socio-economic characteristics and have been found to exert more attention to experimental instructions (Hauser and Schwarz, 2016).[14] In our experiment, subjects that complete the study earn 1.20 USD participation fee plus bonus pay depending on their behavior during the experiment. Tokens constitute the experimental currency at the exchange rate of 1 token=0.005 USD. Completing the experiment took participants 17 minutes and 4 seconds on average. Including participation fee, on average, subjects earned 1.63 USD for themselves, and generated 1.13 USD donations for the charity of our choice – *Médecins Sans Frontières*. For subjects that do not spend time on the donation task, the experiment only took 10 minutes and 33 seconds; these subjects earned 1.34 USD, including participation fee, on average. Such average earnings are comparable to the 7.25 USD hourly earnings accumulated by the most productive 4% of workers on this platform and are significantly higher than the median hourly earnings of 2 USD (Hara et al., 2018). Participation in the experiment is allowed only once, and no retakes are granted to subjects that accidentally drop out of the study.[15]

[13]The experimental software is programmed in oTree (Chen, Schonger, and Wickens, 2016b). We collect data over multiple sessions to minimize risks of overloading our server.

[14]Like other studies conducted on this platform, we restrict participation in our experiment to workers with an approval rate above 90%. We also restrict participation to workers residing in the U.S.

[15]A 40-minute timer is implemented to encourage subjects to complete the experiment timely and without distractions. Furthermore, to discourage speeding behavior and the use of bots, we implement

2.2.4 Randomization Checks

From the total of 3,467 subjects that began the experiment, we work with a sample of 2,914 subjects who completed both the joint problem solving (JPS) task and the donation task. In the JPS task, subjects score an average of 40 out of the 80 available points. After the task, subjects report a 2.8 oneness towards their peer on average (on a scale between 1 and 7). Across the ten treatment conditions, subjects generate 4.6 donations, on average, for *Médecins Sans Frontières*, and predict their peer to generate an average of 3.9 donations. Table 2.1 shows balance in pre-treatment measures and attrition. The lack of differential attrition across treatments attenuates concerns of disappointment effects from our treatment manipulations.

2.2.5 Social Proximity

As argued in the conceptual framework, conformity requires some degree of social connection to the social reference.[16] This section discusses interpretation and determinants of our measure of social proximity, which we elicit among pairs of strangers after they interact in the JPS task.

Recall that in this task, pairs of subjects are presented with four paintings and they need to agree on the correct artist to associate from a list of five artists for each painting. Social contact within each pair occurs in the chat box that allows for instrumental coordination on answers and strategies to solve the task.[17] An average score of 40 out of 80 available points indicates significant coordinated effort to solve the common puzzles; random click-through from both subjects would predict an expected score of 3.2. The chat box also introduces each subject to the peer by reporting peer's stated first name, age, gender, level of experience on the Amazon Mechanical Turk platform, and

a practice of the real effort task before treatment assignment.

[16]Studying behavioral mechanisms that operate via social interactions is methodologically complex. Some papers leverage existing social relationships and identities, while others induce the formation of social relationships and identities within the experiment (Goette, Huffman, and Meier, 2012 and Chen et al., 2014 for reviews of this literature). For our investigation, to avoid contaminating the conformity with other forms of social influence deriving from the prospects of future interactions, we choose the approach of building social relationships among randomly and anonymously matched strangers.

[17]To solve puzzles, many of the subjects realize that they can use Google image search, and they tend to split up paintings to search with their peer.

Table 2.1: Summary Statistics of Observable Characteristics and Attrition(Means and Standard Errors in Parentheses)

Incentives to peer			Zero	Moderate	High	Zero	Moderate	High	Zero	Moderate	High	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Male	0.452	0.449	0.437	0.441	0.465	0.408	0.473	0.451	0.453	0.458	0.484	0.861
	(0.009)	(0.029)	(0.029)	(0.029)	(0.030)	(0.029)	(0.030)	(0.029)	(0.030)	(0.029)	(0.028)	
Age group	2.524	2.491	2.473	2.500	2.620	2.582	2.513	2.487	2.529	2.548	2.500	0.882
	(0.021)	(0.068)	(0.065)	(0.065)	(0.066)	(0.066)	(0.069)	(0.064)	(0.065)	(0.065)	(0.068)	
Experience	2.605	2.774	2.567	2.666	2.662	2.624	2.564	2.632	2.604	2.568	2.403	0.280
	(0.028)	(0.092)	(0.089)	(0.091)	(0.089)	(0.089)	(0.093)	(0.086)	(0.088)	(0.083)	(0.093)	
Points JPS task	40.199	37.979	40.333	40.966	42.324	39.443	41.392	39.934	41.079	39.535	39.226	0.924
	(0.619)	(1.957)	(1.906)	(1.985)	(1.982)	(2.004)	(2.034)	(1.965)	(1.955)	(1.795)	(2.029)	
Oneness	2.801	2.704	2.847	2.784	2.894	2.793	2.885	2.773	2.831	2.691	2.819	0.829
	(0.030)	(0.093)	(0.096)	(0.097)	(0.097)	(0.098)	(0.103)	(0.094)	(0.095)	(0.096)	(0.089)	
Dropout	0.159	0.138	0.167	0.167	0.147	0.171	0.152	0.163	0.165	0.169	0.151	0.974
	(0.006)	(0.019)	(0.020)	(0.020)	(0.019)	(0.020)	(0.020)	(0.019)	(0.020)	(0.019)	(0.020)	
Observations	2914 [3467]	287 [333]	300 [363]	290 [348]	284 [333]	287 [346]	273 [322]	304 [363]	278 [333]	301 [362]	310 [365]	

Notes: p-value in column (12) is for a one-way ANOVA on ranks (Kruskal-Wallis) test comparing the ten treatment groups in columns (2) to (11). Except for dropout rates ("Dropout"), all statistics refer to the final sample of subjects who completed the experiment. Dropout rates of subjects after treatment assignment computed on the samples reported in square brackets in the "Observations" row.

common US residence. The *oneness* measure of social proximity is meant to capture the extent to which basic demographic information and contact with the other player in the JPS task facilitate the formation of perceived social proximity.[18]

To put into perspective the kind of social proximity captured by the oneness scale, it is worth comparing the levels we measure to existing estimates. In other studies, on the same scale from 1 to 7, oneness towards an acquaintance, non-close friend, and close relationship is measured to be on average 2.5, 4.0, and 5.4, respectively (Gächter, Starmer, and Tufano, 2015). In our sample, we measure greatly different levels of oneness, with an inter-quartile range capturing half of the entire range of possible realizations: the first quartile of the distribution is 1, the median is 2.5, the third quartile is 4. Expectedly, many subjects exhibit no social proximity to their peer in the experiment. But it is interesting to notice that at least half of the sample exhibits social proximity towards their peer – a stranger with whom they have recently made contact to solve puzzles – similar to social proximity that other studies observe towards acquaintances. This is not a *causal* effect of JPS interactions on social proximity, but gives an indication that the JPS does harness social proximity. More direct causal evidence can be found in Gioia, 2017.

In **??**, least squares regressions illustrate the correlates of social proximity, and highlights the role of both *homophily* (Marmaros and Sacerdote, 2006) and chat box contact (Chen and Li, 2009) in the formation of social proximity. Although age difference between the paired players does not seem to be highly predictive of social proximity, the peer being of the same gender and having similar experience on the platform predict significantly higher oneness. The fit of this simple linear regression model improves remarkably when we include a binary indicator – *contact* – for whether players made reciprocal contact through the chat box provided.[19] Players that make reciprocal con-

[18]Figure B.4 provides the distribution of the two psychological scales underlying oneness. These two scales are strongly correlated ($\rho = 0.731$), with the WE scale exhibiting a relatively multi-peaked distribution compared to the clear single peak of the IOS scale (at the lowest level of social proximity). All analyses presented in the results section are robust to replacing either of these two scales as measures of social proximity.

[19]80.4% of subjects used the chat box to make contact with the peer, and 64.6% of pairs managed to have a conversation (*contact* = 1). In these conversations, subject share their knowledge of the paintings, share relevant personal information and considerations (e.g., one says "If my husband was here he

tact with their peer report 67.5% higher social proximity, and although the decision to engage in chat interactions is endogenous, the relatively strong correlation of 0.294 (column (1)) is indicative of the role of social contact for the development of social connection.

2.3 Experimental Results

2.3.1 Descriptive Evidence

In Table 2.2, we summarize average beliefs and donations across treatments, drawing the patterns of interest that will be explained in the next section.

Own incentives. Donations are weakly monotonic in personal incentives. They strongly increase when incentives go from *zero* to *moderate* and appear to flatten out when incentives are *high*. *Moderate* incentives also increase donations compared to a control treatment in which subjects are not incentivized and incentives are never mentioned. Beliefs indicate that individuals anticipate these patterns of direct incentive effects correctly, although they systematically underestimate the levels of others' generosity.[20]

Peer incentives. Donations systematically respond non-monotonically to peer incentives: for any level of personal incentives, donations increase when peer incentives go from *zero* to *moderate*, and decrease when incentives go from *moderate* to *high*. Subjects anticipate such comparative statics remarkably well. They anticipate that an increase in their own incentives from *zero* to *moderate* is going to increase donations of their peers and a further increase from *moderate* to *high* decreases their donations.

2.3.2 Evidence of Social Influence

In this subsection we test the statistical significance of these patterns and interpret the evidence through the lens of our social influence framework. We also examine

would know, he is an art teacher lol", some other says that "Modern art sucks".), and agree upon strategies to solve the task (e.g. "You betcha. I'm googling the heck out of it right now. I've got Miro for the first one, Botticelli for the second, Grant Wood for the 3rd, working on the 4th."). Scripts of these conversations can be made available upon request.

[20]Consistent with studies finding that research subject accurately predict experimental results (DellaVigna and Pope, 2016), but people underestimate others' prosocial attitudes (e.g. Goette, Huffman, and Meier, 2006).

Table 2.2: Beliefs and Donations Across Treatments (Means and Standard Errors)

		Beliefs about peer's donations			Own donations		
Incentives offered							
No (control)		3.585			3.934		
		(0.205)			(0.222)		
Yes (3x3 treatments)							
		Incentives to peer			Incentives to peer		
		Zero	Moderate	High	Zero	Moderate	High
Incentives to self	Zero	2.540	4.331	4.637	3.233	3.417	3.190
		(0.182)	(0.215)	(0.208)	(0.217)	(0.230)	(0.210)
	Moderate	2.585	4.832	5.086	5.042	5.546	5.155
		(0.193)	(0.213)	(0.207)	(0.233)	(0.235)	(0.224)
	High	2.374	4.100	4.374	5.299	5.575	5.187
		(0.174)	(0.201)	(0.195)	(0.233)	(0.229)	(0.212)

whether the evidence can be explained by other theories of prosocial behavior.

$$Donation_i = \alpha + \beta_1 Lottery_i + \beta_2 Moderate_i + \beta_3 High_i + \beta_4 Moderate_j + \beta_5 High_j + \boldsymbol{X}_{i,j}\boldsymbol{\gamma} + \varepsilon_i \tag{2.4}$$

We use a linear regression model (2.4) to estimate how donations are affected by the economic environment. Denoting an agent by i and her peer by j, this model estimates both the effect of i's incentives on i's donations as well as the effect of j's incentives on i's donations. We allow for the effects of incentives to be non-monotonic by coding incentive as categorical variables. The regression model also includes an indicator for the *no lottery* control treatment that isolate disappointment effects of not receiving the incentives, as well as controls for observable characteristics of both players in a pair.

$$Belief_i = \phi + \delta_1 Lottery_j + \delta_2 Moderate_j + \delta_3 High_j + \delta_4 Moderate_i + \delta_5 High_i + \boldsymbol{X}_{i,j}\boldsymbol{\omega} + \epsilon_i \tag{2.5}$$

We also estimate the mirror regression model (2.5) for individual beliefs on the donations of her peer. This allows us to test the unique prediction of the social influence framework that changes in peer incentives can cause a change in beliefs about how peer j behaves and a change in the behavior of agent i in the same direction. The estimates of regression models (2.4) and (2.5) are presented in panels (a) and (b) of

Table 2.3, respectively.

Consistent with the descriptive evidence, column (1) of panel (a) shows that increasing an agent's incentives from *zero* to *moderate* increases donations by 1.964 units ($p < 0.001$), while increasing incentives from *moderate* to *high* does not lead to a further increase in donations ($p = 0.649$). Agents also respond to changes in peer incentives, but do so non-monotonically (column (2)). Increasing peer incentives from *zero* to *moderate* increases donations by 0.356 units ($p = 0.055$), but donations drop when the peer incentives further increase from *moderate* to *high* ($p = 0.046$).

Panel (b) shows that agents anticipate these incentive effects. They anticipate that increasing someone's private incentives from *zero* to *moderate* will have a strong positive effect, and the effect of increasing incentives further will be subtle. They also predict that their peers will react to peer incentives non-monotonically. In fact, they believe that an increase in their own incentives from *zero* to *moderate* causes their peer to donate 0.336 extra units ($p = 0.044$), but a further increase in their own incentives from *moderate* to *high* would cause peer donations to drop ($p < 0.001$).

The non-monotonicity of donations in peer incentives is driven by subjects with a strong connection to their peer. When we estimate (2.4) and (2.5) separately for subjects above and below the median level of social proximity we find that socially distant subjects monotonically increase donations with monetary incentives, and they expect their peer to do the same. Yet, their giving behavior is not significantly affected by the incentives provided to their peer.[21] If at all, monetary incentives to the peer monotonically decrease a subject's own giving: donations decrease by 0.214 units and 0.251 units when the peer gets moderate and high incentives, respectively. Notwithstanding, these point estimates are not significantly different from zero. Socially close subjects respond differently to changes in the incentives of their peer ($p = 0.016$ for joint F-test for equality of effects between high and low oneness subjects).[22] When

[21] We partition the sample at the median score of oneness. For robustness, we try sample splits at the median score of the JPS task and at the median score of just one of the two psychological scales that are used to construct oneness; the results are qualitatively the same.

[22] In Table B.2.4 we illustrate the robustness of this result in a pooled specification that interacts the treatment with an indicator for high social proximity.

Table 2.3: Incentive Effects on Donations and Beliefs

		Full sample		Split by oneness		H_0 p-value:
				High	Low	*High = Low*
Outcome:		(1)	(2)	(3)	(4)	(5)
(a) Donations						
	Provided Lottery	-0.712***	-0.831***	-0.665*	-1.066***	0.464
		(0.262)	(0.283)	(0.389)	(0.403)	
	Incentives to self *(baseline: Zero)*					
	Moderate	1.964***	1.970***	1.921***	2.037***	0.052
		(0.183)	(0.182)	(0.254)	(0.260)	
	High	2.047***	2.044***	1.712***	2.502***	
		(0.179)	(0.179)	(0.242)	(0.259)	
	Incentives to peer *(baseline: Zero)*					
	Moderate		0.356*	0.837***	-0.214	0.016
			(0.186)	(0.259)	(0.268)	
	High		-0.001	0.170	-0.251	
			(0.180)	(0.236)	(0.269)	
	Constant	4.663***	4.650***	4.896***	4.248***	0.369
		(0.368)	(0.368)	(0.500)	(0.539)	
	Incentives to self, *High - Moderate*	0.083	0.073	-0.209	0.465*	
		(0.182)	(0.182)	(0.246)	(0.272)	
	Incentives to peer, *High - Moderate*		-0.356**	-0.667***	-0.037	
			(0.178)	(0.245)	(0.262)	
	H_0 p-value: Incentives to peer					
	Zero = Moderate = High = 0		0.080	0.003	0.607	
(b) Beliefs						
	Provided Lottery	-1.155***	-1.188***	-1.207***	-1.315***	0.822
		(0.237)	(0.256)	(0.358)	(0.348)	
	Incentives to peer *(baseline: Zero)*					
	Moderate	1.948***	1.962***	2.155***	1.773***	0.391
		(0.161)	(0.160)	(0.222)	(0.221)	
	High	2.237***	2.240***	2.227***	2.218***	
		(0.158)	(0.158)	(0.211)	(0.229)	
	Incentives to self *(baseline: Zero)*					
	Moderate		0.336**	0.420*	0.257	0.435
			(0.167)	(0.222)	(0.240)	
	High		-0.253	-0.337	-0.105	
			(0.160)	(0.221)	(0.227)	
	Constant	4.273***	4.274***	4.800***	3.625***	0.075
		(0.341)	(0.341)	(0.458)	(0.495)	
	Incentives to peer, *High - Moderate*	0.288*	0.278*	0.072	0.445*	
		(0.168)	(0.167)	(0.226)	(0.242)	
	Incentives to self, *High - Moderate*		-0.589***	-0.757***	-0.362	
			(0.158)	(0.219)	(0.225)	
	H_0 p-value: Incentives to self					
	Zero = Moderate = High = 0		0.001	0.003	0.267	
	Observations	2914	2914	1571	1343	

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: All specifications include gender, age group, and experience, of both the player and her peer, as well as session dummies. Column (5) presents joint F-tests for the null hypotheses that point estimates – for each group of variables – are equal in the high and low oneness subsamples. Standard errors are clustered at the pair level. Results are qualitatively very similar in a seemingly unrelated regression framework that allows for correlation in the error term of individual beliefs and donations.

peer incentives increase from zero to moderate, subjects expect their peer to increase donations by 2.155 ($p < 0.001$) units and they donate 0.837 ($p < 0.001$) units more themselves. However, when peer incentives increase from *moderate* to *high*, subjects again believe that the incentive increase does not affect ($p = 0.750$) peer donations (correctly so given that such increase in incentives for high oneness subjects does not increase donations significantly ($p = 0.395$)), and donations *decrease* by 0.667 ($p = 0.007$) units and individuals.[23]

The evidence is clear that the economic environment of the peer shapes willingness to behave prosocially. This is evidence of conformity, with *zero*-to-*moderate* changes in incentives causing individual donations and beliefs about the donations of others to move in the same direction. At the same time, the evidence indicates that the desire to conform diminishes when peer incentives are "too generous". Fewer donations are made when peer incentives are *high* in spite of the fact that expected donations from the peer do not decrease.

This evidence is explained by a model of social influence in which conformity is driven by identification with altruistic intentions. As we show in Section 2.2.2, such model captures that changes in peer incentives not only affect donations of the peer, but also the intensity of their altruistic intentions. The effects of peer incentives on donations are non-monotonic because any loss in psychological utility due to norm deviation is dampened when the norm is determined by weak altruistic intentions.

Wald estimates for the effect of beliefs on donations implied by our reduced form regressions results help appreciate diminishing conformity more directly. From column (2), we obtain that a one unit change in beliefs from increasing the peer's incentives from *zero* to *moderate* increases donations by 0.182 units ($p = 0.035$), while a one unit change in beliefs from increasing the peer's incentives from *zero* to *high* has no effect on donations ($b_{Wald} = -0.0003$, $p = 0.997$). This pattern of diminishing influence of beliefs about others on individual prosocial behavior is even more pronounced for

[23]Importantly, differences in behavior across socially close and socially distant individuals does not appear to be driven by differences in pro social orientation. In fact, we can use the control treatment to show that in the absence of incentives subjects with high social proximity to their peer do not systematically donate differently from subjects with low social proximity to their peer ($p = 0.973$, see Figure B.5).

high oneness subjects (column (3)).[24]

Alternatively, the non-monotonic effects of peer incentives on donations may be driven by a substitution effect due to altruistic crowding-out. Altruistic agents may decrease their donations when they expect that incentives cause the peer to increase donations. However, in such a model, an agent's donations decrease globally with her peer's incentives: the reason is that increased donations by the peer always lower the marginal utility of one's own donation. Thus, impure altruism alone cannot explain our findings, as we find non-monotonic effect of the peer's incentives on the agent's behavior.

Another hypothesis is that substitution effects co-exist with conformity and explain the diminishing conformity when peer incentives are *high*. This hypothesis is also at odds with the evidence. Low oneness subjects believe that their peer respond to incentives strongly and monotonically, but their donations do not respond to the incentives of their peer ($p = 0.607$). All the non-monotonic response to peer incentives is driven by high oneness subjects. However, the pattern in this group is also inconsistent with the substitution hypothesis. They believe that changing incentives for their peer from *moderate* to *high* has no significant effects on the peer's donations ($p = 0.750$) and yet they react by reducing their own donations by 0.667 units ($p = 0.007$).

The diminishing conformity interpretation is reminiscent of influential papers by Gneezy and Rustichini, 2000a; Gneezy and Rustichini, 2000b and the more recent study of Fuster and Meier, 2009. From their experiments, these authors conclude that incentives weaken adherence to the norms of behavior dictated by the actions of a social reference – let this be a small group or society. An important novel element of distinction of our findings is that incentives do not seem to simply shut down adherence to social norms: in fact, the magnitude of incentives matters. Relatively small incentives to act prosocially can preserve a certain level of norm adherence and produce social influence.[25] When this is the case, our evidence suggests that larger

[24]The belief change due to increasing the peer's incentives from *zero* to *moderate* increases one's donations by 0.388 units ($p < 0.001$), while the belief change due to increasing the peer's incentives from *zero* to *high* has a precise null effect on donations of 0.076 ($p = 0.448$).

[25]Ostracism as in Dutta, Levine, and Modica, 2018 allows us to endogenize social norms to demon-

incentives are more likely to backfire on the positive spillovers of social influence, and perhaps the power of small (but not large) incentives could be leveraged to crowd in donations.

2.3.3 Incentive Inequality and Donor's Morale

In the interpretation of our results, we have so far ignored the possibility that incentive inequality in itself may affect an agent's morale to work on a task to generate donations for a charity. To assess this potential mechanism, we consider a model that incorporates such effects from incentive inequality (Breza, Kaur, and Shamdasani, 2017). Such a model predicts that conditional on one's own incentives, donation levels should be highest when incentives for both players in a pair are equal, and monotonically decrease with the gap between one's own and peer's incentives. In our setting, this implies a set of inequality relationships in average level of giving between treatments, that we derive in Section B.1.3 and summarize in Table 2.4. We refer to these inequality relationships as the *main diagonal condition*.

Table 2.4: Inequalities in Average Donations between Incentivized Treatments Predicted by the Main Diagonal Condition

		Incentives to peer				
		Zero		Moderate		High
Incentives to self	Zero	$\mu_{n,n}$	$>$	$\mu_{n,m}$	$>$	$\mu_{n,h}$
	Moderate	$\mu_{m,n}$	$<$	$\mu_{m,m}$	$>$	$\mu_{m,h}$
	High	$\mu_{h,n}$	$<$	$\mu_{h,m}$	$<$	$\mu_{h,h}$

These predictions immediately appear in contrast with raw averages of donations presented in Table 2.2, where we observe that conditional on the agent's private incentives, increasing inequality often leads to more donations. Instead of focusing on local violations, we devise a likelihood ratio test of the joint null hypothesis that the *main diagonal condition* explains the first moments of donations and beliefs in our data (Burks et al., 2009). These tests, which are discussed in detail in Section B.2.1, largely reject the

strate that it is not the mere incidence of payments that damages norm following, but sufficiently large incentives are instead needed. Albeit aligned with our evidence, for the absence of social interactions *after* the donation, we cannot meaningfully use this theory to explain our findings.

joint hypothesis for both donations and beliefs. Rejections are especially strong when we focus on the behavior and beliefs of high oneness subjects. Taken together these findings rule out incentive inequality as an explanation for our social influence effects.

2.3.4 Other Mechanisms of Social Influence

Mechanisms such as social learning, social consumption, reciprocity, and conformism have been proposed to explain the large body of evidence in support of the hypothesis that most individuals are conditional co-operators (Frey and Meier, 2004). In this section we discuss other mechanisms that can generate spillovers of giving in applications similar to the one we consider. Further, we explain how the experimental design allows us to rule out these explanations.

Social learning. When people are asymmetrically informed about relevant parameters, observing others' behavior can facilitate information aggregation. In any charitable giving context, the social value of a prosocial activity is uncertain, and the attitudes of others towards the charitable activity may indeed be informative about the quality of the charity or the social norm of giving to the specific cause. Our experiment excludes any scope for social learning. We make clear to subjects that the value generated from a donation is 0.25 USD and that this is common knowledge. Yet, the effectiveness of *Médecins Sans Frontières* in generating social value may be uncertain and some subjects may know the charity better than others. Our experiment rules out this channel by keeping donations private.

Joint consumption. Especially for volunteer work, this mechanism can play an important role in producing social influence. The prosocial action may involve social activities that confer consumption utility from forming relationships, sharing common experiences, and other pleasant interactions during the activity. The lack of social interactions among participants during the donation makes it easy to rule out this mechanism.

Reciprocity. This mechanism is often used to explain behavior in *local* social dilemmas - where agents directly benefit from the prosocial behavior of others. In most cases,

charitable giving can instead be regarded as a *global* social dilemma, in the sense that agents only benefit marginally from the prosocial behavior of others. In such settings, we cannot expect that reciprocity could generate first order effects.

Signaling motives. The theory of Benabou and Tirole, 2006 proposes the signaling of altruism and greed as channels that endogenuously lead to strategic complementarity or substitutability of donations. They show that complementarities arise when, as more people decide to donate, the image of the pool of donors deteriorates faster than the image of non-donors. Our context is highly anonymous and our results are unlikely to be driven by *social* signaling. At the same time, we recognize that the Benabou and Tirole, 2006 model admits a self-image interpretation.[26] However, for self-signaling to explain variation in donations, the treatment should affect the inference individuals can make about their own type, which cannot be in our setting where peer incentives are random.[27]

Social influence in work effort. One possibility is that the social influence observed in this study may have to do with conformity in work effort rather than in prosocial behavior. While we do not have a parallel experiment to rule out this channel, prominent existing studies on social influence in the workplace (e.g. Mas and Moretti, 2009; Bandiera, Barankay, and Rasul, 2010) show that some degree of socialization or observability of co-workers' effort during the activity is necessary for this channel to matter empirically. Because subjects in our experiment work on the real effort task for the charity in isolation from their peer, we believe that this channel plays a trivial role, if any.

2.4 Conclusion

This study proposes a novel experiment to study social influence independently of social learning. In our experiment, pairs of players collaborate on a task that provides the opportunity to develop social proximity with their peer. Each individual then

[26] Especially, we do not dispute that signaling motives and conformity may have related behavioral roots. Jones and Linardi, 2014 find that making signaling motives more salient increases conformism.

[27] Random assignment of peer incentive m_j implies that for inference about own altruism type v_i, without observing peer donation d_j, $E_i(v_i|d_i, m_i, m_j) = E(v_i|d_i, m_i)$.

independently generates donations to a charity through a tedious task, knowing both her incentives and the incentives of her peer.

We provide evidence that information about the economic environment faced by a social reference is sufficient to spread social influence. Agents respond to increases in their peer's incentives by expecting that their peer will donate more and in turn, they donate more themselves. Our result are in line with a model of social influence in which conformity is driven by identification with an attractive role (Kelman, 1961). We find that conformity in donations is stronger when the agent feels socially close to her peer, and her response to the incentives of the peer are non-monotonic.[28]

Our results also have methodological implications. Increasingly, social scientists are becoming interested in studying the relationship between beliefs about others' behavior and individual behavior. Such empirical efforts often have to overcome several challenges, including the notorious reverse causality issue of *false consensus*.[29] An approach that is increasingly used in experiments to overcome similar challenges is to introduce sources of belief variation that serve as instruments for beliefs (see e.g. Smith, 2013; Costa-Gomes, Huck, and Weizsäcker, 2014). The non-monotonicity of donations in peer's incentives is a warning that different incentives can generate IV estimates that are potentially contradictory if we do not account for the model through which beliefs cause behavior.

This evidence is informative of the mechanisms underlying conformity. As noted by Dutta, Levine, and Modica, 2018, whether conformity is a preference or a social norm is difficult to say in most empirical settings. An individual may prefer to *internalize* social norms instead of doing the introspection needed to determine her favorite strategies. While we do not make this distinction, we think that our design makes it hard for individuals to internalize social norms for these not being readily available. In fact, because others' behavior is not observable, in order to enjoy any of the bene-

[28]Our setup does not distinguish between probabilistic and deterministic incentives. While it is possible that this probabilistic framing reinforces the effects, it is difficult to see how the framing alone (without the higher expected payment) would generate the non-monotonicity we observe.

[29]The concern that beliefs reflect more the response function of the *observer* than the one of the *observed*.

fits of internalization, subjects would first have to accurately assess what is the social norm in a relatively unfamiliar environment.

An implication of our results is that small incentives can be more effective at crowding in prosociality, and non-pecuniary interventions may be better suited to leverage social influence. Market designers should be cautious with changing incentives for activities that are partly regulated by a social contract because larger incentives are more likely to backfire on social influence. Consistent with this interpretation is the surprising evidence that *better* paid police officers in West Africa become *more* corrupt (Foltz and Opoku-Agyemang, 2015).

More broadly, by distinguishing conformity from social learning, our results illustrate the potential of social influence even in settings where social information is unlikely to be informative of the quality of an activity. This improves our understanding of the propagation of social influence in other applications, like exercising (Aral and Nicolaides, 2017) and political mobilization (Bond et al., 2012), water and energy conservation (Ferraro and Price, 2013; Allcott and Rogers, 2014) where personal tastes are likely unaffected by social information. A concern in this literature is that social-norm information can be a double-edged sword, for it may lead to bunching of outcomes around the norm. However, many of these recent field studies find that social-norm information also leads to adjustment in the socially desirable direction for individuals that are already doing better than what is dictated by the social norm. Our results suggest that previous findings can be explained by conformity to expectation of how others will react to social information. Assessing the portability of our results is an interesting avenue for future research.

Chapter 3

Sorting Into Incentives for Prosocial Behavior

3.1 Introduction

Many public goods rely on voluntary private contributions. Millions of people every year spend their time working as volunteers in their communities, give money to charity, or donate their own blood, organs, and other tissue. For charities seeking volunteers or money and for health care providers seeking blood donations, it is important to understand how to encourage this prosocial behavior.

An often-used way is to provide extrinsic incentives. The economics literature has found mixed evidence on the effects of monetary and non-monetary incentives on giving (Frey and Oberholzer-Gee, 1997; Bowles and Polanía-Reyes, 2012). Although a positive effect of extrinsic incentives is in line with standard economic theory, it goes against a considerable literature in psychology and economics, which argues that they can backfire by either crowding out the intrinsic motivation to give (Deci, 1975; Deci, 1971; Titmuss, 1971), or ruining the reputation of donors who could be regarded as greedy (Benabou and Tirole, 2006; Exley, 2017). Field experiments have found evidence for extrinsic incentives to have both negative effects on volunteer work (Frey and Goette, 1999) as well as positive effects on organ (Lacetera, Macis, and Stith, 2014) and blood donations (Lacetera, Macis, and Slonim, 2012; Lacetera, Macis, and Slonim, 2014).[1]

While the role of incentives has been analyzed in a wide range of domains, they have been mostly studied in isolation and contrasted to the absence of incentives. In this paper, we study a setting where different incentives coexist. In this environment, agents can turn down an extrinsic incentive to donate. This lets them reveal and signal

[1] Aside from the question of effectiveness, incentives to donate human tissue might be seen as controversial on moral grounds. Only limited incentives appear to be morally acceptable among a sample of people surveyed in the United States (Boulware et al., 2006). Becker and Elias, 2007 provide a compelling argument in favor of allowing incentives for organ donations. Lacetera, 2016 summarizes the debate. In this paper, we will abstract from the matter of the morality of incentives.

their individual preferences through their actions.

Our setting is motivated by the market for human whole blood donations in Germany.[2] In most high-income countries, the concern that incentives could backfire is reflected in tight regulation of how blood donations can be collected. Regulations typically do not allow for monetary payments to donors (The Lancet, 2005; World Health Organization, 2009; Council of Europe, 1995). In many regions of Germany, however, monetary and non-monetary incentives appear to coexist in a "dual market" in which different blood collectors offer different incentives and prospective donors can choose where to donate. Donations at the Red Cross are always unpaid, while donations at hospitals or commercial blood banks are compensated with 20 to 30 euro.

Very little is known about the features of such "dual markets" for the collection of charitable contributions. Does this system of collection increase donations compared to a single market in which either everyone is unpaid or everyone is paid? What are the determinants of the share of unpaid donations in a dual market? In this paper we focus on two channels that could help explain sorting into unpaid donations in a dual collection system: altruism and social image concerns.

To guide our analysis, we use a model of charitable giving in which prospective donors are motivated to give by intrinsic incentives, extrinsic incentives, and image concerns. We build on the framework by Benabou and Tirole, 2006, but introduce two modifications: first, we change the payoff structure so that a potential compensation for the donation is paid out of the value that is generated by the donation. This tension between private and public benefit of the donation introduces a channel through which extrinsic incentives can crowd out intrinsic motivation. Second, we assume that donors do not differ in how much they value extrinsic incentives. This lets us make clear predictions, but comes at the cost of ruling out "reputational crowding out", that

[2]The most common type of human blood donation is a "whole blood" donation, in which approximately one pint of blood is collected over a period of about ten minutes. Men can donate up to six times per year, women up to four times per year. Red blood cells from whole blood donations are typically used for transfusions to other patients and are most commonly seen as motivated by altruistic preferences (Niessen-Ruenzi, Weber, and Becker, 2015). Other types of blood donations include platelet and plasma donations, which take much longer and require donors to be connected to a machine. Donors are commonly compensated in cash for these types of donations.

is we rule out that donors can have a negative response to the introduction of extrinsic incentives out of concern for appearing greedy. We derive three testable behavioral hypotheses from this model.

The first testable hypothesis states that the availability of compensation to donate should increase donations. We call this the "incentive effect". The second hypothesis states that irrespective of whether compensation is available, making actions visible should increase donations. We call this the "social image effect". Our third and novel hypothesis states that in a dual market, where agents can turn down compensation, a positive share of agents will choose to remain unpaid and that this share is larger when actions are taken in public. We call this "sorting", based on the idea that a dual market can bring about efficiency gains in the collection similar to those deriving from self-selection in second-degree price discrimination.

We test these three hypotheses in a laboratory experiment with 329 student subjects. For three rounds, each subject is confronted with the decision to participate in a real effort task. This task generates value for a charity under one of three market designs: donors receive no compensation for the donation (single market *NOT PAID*), donors always receive a compensation for the donation (single market *PAID*), and donors can choose whether they want to receive compensation for the donation (dual market *CHOOSE*). Like for the case of blood collection, any compensation paid out to donors reduces the social value of the donation. This is objectively measured in our controlled setup by the amount of money that goes to charity. We also vary the visibility of actions (*PRIVATE* vs. *PUBLIC*). The combination of market design treatments and visibility treatments in a full 3×2 design produces six distinct treatments, which we run between subjects.

The experimental results mostly support our behavioral hypotheses. We find clear evidence for the incentive effect. In the dual market, the availability of incentives does not crowd out intrinsic motivations of donors, irrespective of whether actions are observable. Moving from a single unpaid market to a dual market significantly increases the number of donations of our experimental subjects.

We also find evidence of strong social image effects. Making actions observable significantly increases donations in all three incentive schemes. Finally, we find support of our sorting hypothesis: when given the option to turn down compensation, a significant share of donors chooses to do so, though we do not find a significant difference between actions taken in private and in public.

Interestingly, and in contrast to similar studies that analyze the effectiveness of conditional and unconditional incentives to act prosocially (Ariely, Bracha, and Meier, 2009; Carpenter and Myers, 2010), we do not find that social image effects attenuate incentive effects. We differ from Ariely, Bracha, and Meier, 2009 in that subjects decide to donate in the presence of an outside option. Our results suggest that when incentives are small and only partly offset the costs of donating, social image effects and incentive effects need not crowd each other out. In addition, we find heterogeneous effects of social image on contributions that we attribute to gender-specific preferences over signaling. Overall, our findings suggest novel ways to improve mechanisms for the collection of charitable donations by leveraging heterogeneity in individual preferences. Applied to the collection of blood donations, our results may inform the design and regulation of systems that use monetary incentives.

The remainder of the paper is organized as follows: Section 3.2 fixes ideas in a simple theoretical framework and presents testable behavioral hypotheses. Section 3.3 details experimental design and procedures. Section 3.4 presents the results. Section 3.5 concludes with a discussion of the implications of our findings for the market for blood that initially motivated our research.

3.2 Theoretical Framework

In the model by Benabou and Tirole, 2006 (henceforth: BT), being compensated to donate can crowd out donations by spoiling the image of donors. Moreover, any compensation is paid from resources that are exogenous to the economy and is given to donors without affecting the social value of their donation. BT show that whether donors can turn down compensation should not matter, because neither image-indifferent nor

image-concerned agents would want to do so. For image-indifferent agents, it would be a dominated strategy to turn down compensation that does not affect the social value of their donation. Image-concerned agents would be worried that their motivation is questioned: turning down incentives could reveal that they are not acting out of altruism, but just to appear as altruistic while in fact (on average) they are not.

For a dual market like in Germany, where prospective donors can choose from a menu of options, the model would thus predict that no one should turn down compensation. Yet we observe that a considerable share of donors chooses to remain unpaid when they have the choice between donating with a 20 to 30 euro compensation or donating without any compensation. Informational frictions and transportation costs may explain part of this outcome, though these do not appear to be empirically significant.

We suggest that a different payoff structure than the one by BT better fits the case of blood donations and many other charitable activities and could explain why prospective donors would choose to turn down incentives. In our version of the model, any potential compensation for the donation is paid out of the value that is generated by the donation. The collector of donations is a charitable organization that transforms collected donations into social value. To increase donations, the collector may find it optimal to pay donors a dividend from their donation as compensation. Increasing private returns from the donation comes at the expense of the value that the donation generates for the rest of the society. This feature of our setup introduces an additional channel through which incentives could potentially crowd out donations: a crowding out of intrinsic motivation. This channel is consistent with an earlier literature stemming from Deci, 1971; Deci, 1975.

To formulate testable predictions that are directly relevant to our research question, we will substantially simplify the original model by BT. One key simplification is that we assume agents to be homogeneous in their taste for extrinsic incentives. When this is the case, there is no scope for signaling greediness (or a lack thereof). Despite being a common assumption in economics, a potential drawback of making this sim-

plification is that it prevents the reputational crowding out from BT, i.e. a situation where extrinsic incentives reduce the donations of agents who seek to avoid signaling greediness through their actions.

3.2.1 Simple Model

The model economy is characterized by a unit mass of agents indexed by $i = \{1, ..., \infty\}$ and one collector of donations. This economy is analyzed under two different institutional settings. We refer to a *single* market when the collector is bound to pay an exogenously-set compensation $y = \tilde{y} \in \mathcal{R}_+$. We refer to a *dual* market when agents are allowed to choose remuneration $y = \{0, \tilde{y}\}$.

The ***collector*** takes donation d from each agent that decides to contribute and transforms it into social value $B \in \mathcal{R}_+$. For each contribution, the collector pays remuneration $y < B$.

Agents differ along two dimensions: the degree of altruism $a_i \sim F(\cdot)$ with positive bounded support, and the concern for image x_i, which we treat as binary with x_i taking value 1 with probability q (and 0 with probability $1 - q$). Both a_i and x_i are independently distributed random variables. Agents make a decision to contribute $d = \{0, 1\}$ in exchange for remuneration y while facing a private cost c. Image concern matters for agents when actions are taken in public ($v = 1$) and is irrelevant when actions are taken in private ($v = 0$).

The utility of agent i can be written as follows:

$$U_i(d, y) = (1 - vx_i)[a_i(B - y) + y - c]d + vx_i E(a|d, y) \tag{3.1}$$

where $E(a|d, y)$ is the image that other agents have of agent i given her actions.

From this theoretical setting we derive two propositions that underpin our analysis:

Proposition 1 (Price discrimination)**.** *A dual market for donations increases contributions compared to a single market where no compensation is available. Compared to a single mar-*

ket where compensation cannot be turned down, allowing agents to turn down compensation reduces the cost of collection without affecting the number of donations.

Proof in Appendix A.

The proposition characterizes the effect of various compensation schemes on donations. It applies when actions are taken in private and in public. Introducing extrinsic incentives to donate increases donations, irrespective of whether these incentives can be turned down. Allowing people to turn down incentives, introduces another margin for people to either express or signal their altruism. Highly altruistic agents donate and choose to turn down the compensation.

As a result, when incentives can be turned down, average cost of collection decreases without compromising supply of donations. These two results illustrate how a dual market, where agents are allowed to choose a remuneration, can bring about efficiency gains in the collection similar to those deriving from self-selection in second-degree price discrimination.

The following proposition is directly linked to the previous and highlights the interaction of image effects with price discrimination.

Proposition 2 (Image effect)**.** *The visibility of actions (i) increases participation in the single as well as in the dual market, and (ii) lowers the average cost of collection in the dual market.*

The proof of (ii) follows directly from the observation that the objective of image-concerned agents who are sufficiently altruistic to donate in private, but not altruistic enough to turn down compensation $y = \tilde{y}$, changes when acting in public. In order to improve their social image, these agents want to pool with the most altruistic agents, who turn down incentives.[3] Part (i) is due to the fact that image-concerned agents only care about their image when acting in public. As a result, even the least-altruistic of these decide to contribute in public in order to avoid the stigma of looking like the selfish segment of the population.

[3]This signaling game may not have an equilibrium in pure strategy if the share of image-indifferent agents who are altruistic enough to turn down the incentives is positive but small compared to the share of image-concerned agents.

3.2.2 Behavioral Hypotheses

We re-organize the predictions contained in the two propositions above into three testable hypotheses. The *incentive effect* and *social image effect* hypotheses immediately derive from propositions 1 and 2, respectively. The *sorting* hypothesis consolidates predictions from both proposition to summarize the interaction of social image effects and incentive effects in the dual market for charitable giving.

Hypothesis 1 (Incentive Effect)**.** *Irrespective of whether actions are visible, the availability of incentives increases donations.*

Hypothesis 2 (Social Image Effect)**.** *Irrespective of whether compensation is available, making actions visible increases donations.*

Hypothesis 3 (Sorting)**.** *In a dual market, a positive share of agents chooses to be not paid. This share is larger when actions are taken in public.*

The incentive effect is consistent with an empirical literature on incentives for donating blood (Mellstrom and Johannesson, 2008; Lacetera, Macis, and Slonim, 2012; Lacetera, Macis, and Slonim, 2013; Niessen-Ruenzi, Weber, and Becker, 2015). Maybe most closely related to ours is the work by Mellstrom and Johannesson, 2008, who conduct an experiment that offers monetary payments to prospective blood donors. Their findings suggest that for women (but not for men), monetary incentives can lead to a net crowding out of donations – thought it is difficult to say whether the results are driven by social signaling or by the fact that incentives lead to a shift in the perception of the incomplete contract, similar to the finding of Gneezy and Rustichini, 2000b. Moreover, they find that letting women turn down the compensation in favor of a donation to charity fully counteracts this crowding out. Our theoretical setup can partly explain this counteracting effect, in that for the most altruistic donors ($a_i > 1$) introducing incentives for charitable giving causes a net utility loss. Such utility loss can be undone when incentives can be turned down in the dual market. In a related paper, Chao, 2017 suggests that even opt-in gifts could crowd out donations if

they shift attention away from the intrinsic motivation. In our framework, we abstract from attention as a potential channel for crowding out.

The social image effect is consistent with a growing empirical literature on the effect of social image or social pressure on charitable actions in particular and economic behavior more generally (Ariely, Bracha, and Meier, 2009; Carpenter and Myers, 2010; Filiz-Ozbay and Ozbay, 2014; Lacetera and Macis, 2010; Bursztyn and Jensen, 2017). Our theoretical setup predicts that, no matter the incentive scheme, making actions visible should increase donations. Consistent with our prediction, Landry et al., 2006 find that both when a charity donation entitles to a lottery ticket and when it does not, social image concerns do increase monetary donations in a door-to-door fundraiser. They also find pronounced gender differences, where men are more likely to contribute to a charity when visited by physically attractive female solicitors. The finding that men are more willing to engage in costly signaling of generosity is consistent with costly signaling theory in evolutionary biology (Gintis, Smith, and Bowles, 2001; Smith and Bird, 2000), which posits that prosocial behavior can be instrumental in signaling good character and attractiveness as a potential match. In particular, there is evidence that women in their mating decision place emphasis on signals indicating resource provision (as opposed to just physical attractiveness), which in turn induces men to strategically signal generosity (Eagly and Crowley, 1986; Iredale, Van Vugt, and Dunbar, 2008; Barclay, 2010; Boehm and Regner, 2013). Van Vugt and Iredale, 2013 call men's public good contributions the "human equivalent of a peacock's tail". Although our theoretical setup is silent on gender differences, we are going to investigate these empirically.

Finally, we are not aware of any empirical evidence on the sorting hypothesis as formulated above. It is not obvious whether prospective donors should increase donations when the choice set is augmented in a way to allow signaling of prosocial orientation either through increased donations or by turning down incentives to donate. A large body of evidence on pure and impure altruism suggests that even when donations are completely private, a positive share of prospective donors presented

with the possibility to contribute time and effort – with or without compensation – would choose to donate not paid.[4] Signaling motives should increase the latent utility of acting prosocially. Increasing the visibility of actions could strengthen the signaling motive, potentially increasing the share of unpaid donations. The theory of Benabou and Tirole, 2006 accommodates sorting as described above, but is hard to test empirically. In our theoretical framework, we chose to make substantial simplifications in order to derive testable hypothesis. We take our experiment as a first step to validate this simplified framework and to test simple hypotheses that could guide the field and inform policy on the properties of dual collection systems for charitable donations.

3.3 Experimental Design and Procedures

3.3.1 General Setup

We test our hypotheses in a laboratory experiment. In our experiment, subjects generate value for a charity by participating in a real-effort task. For the experimental task, we build on the "click for charity" design by Ariely, Bracha, and Meier, 2009. Different from Ariely, Bracha, and Meier, 2009, subjects in our framework can choose between participating in the donation task or skipping the task and taking a fixed payoff as outside option.[5] This outside option introduces an homogeneous private cost of donating on top of the individual cost of exerting effort. If subjects choose to participate, they can generate a donation by sequentially entering 400 key sequences on a computer keyboard. One sequence constitutes of four key presses (*"w"*, *"e"*, *"e"*, *"return"*). On their screen, subjects see a bar indicating progress towards the required number of sequences. We chose this task because it is not inherently meaningful or intrinsically rewarding, and allows us to focus on motivation to exert effort for a charity. Other tasks, particularly ones that are more gamified, may be differentially appealing

[4] See Ottoni-Wilhelm, Vesterlund, and Xie, 2014 for a review of the pure and impure altruism literature.

[5] Without the outside option, the marginal cost of participating in the task could be low enough for lab subjects to be indifferent between exerting effort and waiting while others exert effort. The outside option increases the costs of participating in the donation task, so that subjects that are not altruistic and not concerned about social image should not participate in the task – as predicted by the model.

to subjects and thus increase noise and confounds (Charness, Gneezy, and Henderson, 2018). Donations generated with this real-effort task are paid out to a charity chosen by each subject.

We employ a full 3×2 between-subject design where we systematically vary the type of incentives offered to engage in the donation task (*PAID*, *NOT PAID*, *CHOOSE*) and the visibility of actions (*PUBLIC* and *PRIVATE*). Visibility is randomly varied across experimental sessions while the incentives offered are randomly varied across all subjects. Table 3.1 summarizes the design.[6]

Table 3.1: Overview of Treatments

	Not paid $y = 0$	Paid $y = \tilde{y}$	Choose $y \in \{0, \tilde{y}\}$
Private Action $v = 0$	n = 46	n = 48	n = 60
Public Action $v = 1$	n = 47	n = 62	n = 66

Notes: Rows list visibility treatments, columns list incentive treatments. n refers to number of subjects in each treatment cell (total of 329 subjects). y refers to the incentive provided, v to the visibility of actions.

After being assigned to one of six treatments, subjects independently engage in the donation task. After the first round, subjects learn that there will be two more rounds of this task. This lets us test our hypotheses both on the extensive and the intensive margin. Irrespective of the treatment, in each of the three rounds can choose between participating in the donation task or skipping. Throughout the experiment, we use tokens as experimental currency. One token is worth 0.04 euro.

[6]We conducted a pilot study of our experimental design online on Amazon Mechanical Turk ($N = 408$) to inform the choice between a within-subject and a between-subject design. To address concerns that a crowding-out effect of incentives may arise either only in an environment where incentives are introduced as a policy change (within-subject) or only in a market design where people are unaware of alternative institutional environments, we also considered an experimental design that allowed us to study the transition from a single market *NOT PAID* or single market *PAID* market design to a dual market *CHOOSE* market design. In this alternative design, we introduced the dual market to subjects after a first round in the single market design. We did not find evidence that the single market design has any persistent effects. Between- and within-subject designs led to qualitatively similar results. We conclude that the initial treatment has no impact on the effectiveness of the *CHOOSE* treatment. For the current project, we opt for a between-subject design to minimize potential confounders and demand effects (Charness, Gneezy, and Kuhn, 2012). Online Appendix F summarizes the pilot.

3.3.2 Treatments

Along the first dimension of the 3×2 between-subject design we vary the market design, i.e. the availability of incentives to participate in the donation task. In the first two treatments, we either provide monetary incentives to participate in the donation task (single market *PAID* treatment) or no monetary incentives (single market *NOT PAID* treatment). In the third treatment (dual market *CHOOSE* treatment), subjects are presented with both the options of a not paid and a paid donation.

The payoffs are set such that donating generates more value for the charity (100 tokens) than the outside option for the subject (75 tokens). When subjects donate and receive monetary incentives for their donation (50 tokens), those reduce the value to charity (from 100 to 50 tokens). Note that the monetary incentives are always smaller than the outside option. Table 3.2 summarizes the choice set in each of the three treatments and the associated monetary payoffs in tokens.

Table 3.2: Payoffs to Subject and Benefits to Charity, by Treatment and Subject Choice (Experimental Currency: "tokens", 1 token = 0.04 euro)

Treatment	Action space	Payoff to subject	Benefit to charity
NOT PAID	Donate not paid	0	100
	Skip	75	0
PAID	Donate paid	50	50
	Skip	75	0
CHOOSE	Donate not paid	0	100
	Donate paid	50	50
	Skip	75	0

Along the second dimension of the 3×2 between-subject design we vary the visibility of subject actions to make public image salient. In the *PRIVATE* treatment, subjects are informed that their actions will remain anonymous. Subjects are seated at desktop computers separated by divider walls and curtains. To maximize anonymity and to rule out that subjects hear each other type while working on the real-effort task, we play a white noise sound using loudspeakers in the laboratory. We verified that the white noise indeed makes it impossible to hear typing from other workstations. We did not receive any complaints from subjects about this measure. In the *PUBLIC* treat-

ment, before beginning the donation task, we inform subjects that they will be asked to reveal their actions in this task in front of all other subjects in this session. Social image effects thus reflect the full decision environment, including the incentive choice in the dual market *CHOOSE* treatment, that each subject is in. After completing all three rounds we ask subjects to publicly report the number of donations they made.[7] Subjects do so by standing up next to their computer in front of the divider walls. There is no explicit requirement to truthfully report this information.[8] Note, however, that reporting takes place after all decisions have been made.

3.3.3 Procedures

Our theoretical framework asserts that more altruistic individuals are, ceteris paribus, more likely to donate to charity. To check that individual levels of altruism are balanced across treatments, we let all subjects play a simple dictator game before beginning the main experimental task that lets subjects donate to charity.[9] In this dictator game, each subject is randomly and anonymously paired with another subject and chooses to split 20 tokens between herself and the anonymous partner. After testing for subject comprehensions, we let both subjects of the pair play the game as the dictator. At the end of the experiment, the experimental software randomly determines which of the two subjects determines payoffs and the game is resolved.

We then introduce a menu of four charities. Three of those charities are chosen because they are assumed to be well-known among subjects: Doctors Without Borders, the International Committee of the Red Cross, and the World Wildlife Fund. We additionally included the Against Malaria Foundation, which is rated as one of the most effective charities by the independent charity evaluator GiveWell. Subjects are given a

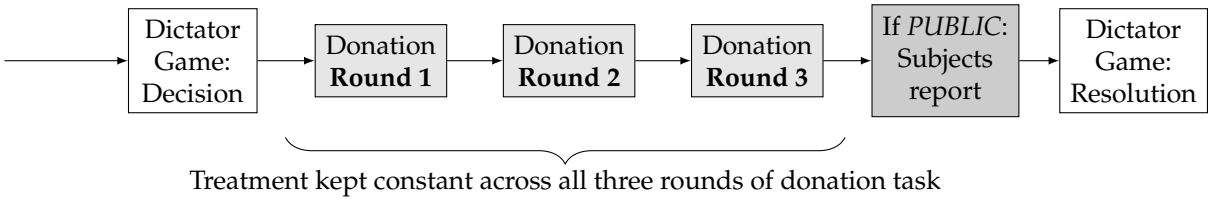
[7] The intention behind not having subjects reveal the incentives under which they donated was to avoid confusion from learning about other treatments.

[8] As an alternative design, we could have publicly announced actual subject choices at the end of the experiment. We decided against this design to stay closer to the theoretical framework of Benabou and Tirole, 2006, where the desire to signal altruism has both instrumental and hedonic origins. We allow for both motivations by letting subjects state their own actions. To maintain the ecological validity of revealing a prosocial action, we do not force subjects to say the truth.

[9] While giving in the dictator game is a well-established measure of generosity vis-à-vis others, it is likely confounded by perceived social norms. As a result, we only rely on our measure of altruism as a balance check, but not to establish key empirical results or to analyze heterogeneous treatment effects.

short description of each charity. We then let each subject choose the charity that they prefer to donate to throughout the experiment. We do this to reduce potential noise from heterogeneous taste for donations to a specific charity. In order to verify balance across treatments, we ask subjects to rate how they perceive each of the charities and how likely they would be to donate money to each of them. Finally, we let subjects practice the donation task before engaging in it for three rounds. In the *PUBLIC* treatment, subjects publicly report their actions after the third round of the donation task. Figure 3.1 summarizes the sequence of tasks in the experiment.

Figure 3.1: Sequence of the Experiment



At the end of the experiment, we collect demographic data. After each session, we confidentially pay out the show-up fee and any earnings that subjects have generated for themselves in the dictator game and the donation task. We also inform subjects about the amount of money donated to charity on their behalf and provide information on how to obtain a confirmation of the donation on their behalf.

We implement the computerized experiment in oTree with our own modifications written in Python and JavaScript (Chen, Schonger, and Wickens, 2016a). A total of 18 experimental sessions were conducted in German at the BonnEconLab in Bonn, Germany, in April 2017 ($n = 329$). Sessions included 20 to 24 subjects and lasted approximately 40 minutes. All subjects are students from various majors at the University of Bonn. They are on average 22 years old, 61% are female. Table 3.3 summarizes the sample. On average, participants earned 10.70 euro for themselves and generated 4 euro for charity.[10]

We can verify that the sample is balanced on observable characteristics, including

[10]Subjects from the pool of the BonnEconLab were invited using hroot (Bock, Baetge, and Nicklisch, 2014). Invitations were restricted to students of the University of Bonn, aged 18–25, with no more than one no-show in prior experiments. Online Appendix B provides further details.

our measure of altruism measured by the dictator game and preference for the chosen charity. Using a nonparametric one-way ANOVA on ranks (Kruskal-Wallis) test, we fail to reject the null hypothesis that the subject pool exhibits the same characteristics across all treatment groups at the 95% level (Table 3.3, column 8).

Table 3.3: Summary Statistics of Observable Characteristics,
Full Sample and by Treatment (Means and Standard Errors in Parentheses)

	Full Sample	Private			Public			p-value
		Not paid	Paid	Choose	Not paid	Paid	Choose	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
a) Measured before treatment								
DG: Tokens kept	15.365	14.891	15.271	15.250	15.021	15.677	15.818	0.848
	(0.214)	(0.621)	(0.558)	(0.507)	(0.618)	(0.501)	(0.411)	
Charity rating	4.602	4.783	4.604	4.583	4.660	4.532	4.515	0.131
	(0.043)	(0.087)	(0.129)	(0.072)	(0.102)	(0.123)	(0.100)	
b) Socioeconomic characteristics, measured after treatment								
Age	21.544	21.630	21.708	21.717	21.511	21.210	21.545	0.499
	(0.091)	(0.263)	(0.223)	(0.213)	(0.263)	(0.184)	(0.207)	
Female	0.611	0.630	0.521	0.717	0.574	0.613	0.591	0.429
	(0.027)	(0.072)	(0.073)	(0.059)	(0.073)	(0.062)	(0.061)	
College major	4.398	4.239	4.417	4.400	4.383	4.661	4.258	0.814
	(0.100)	(0.277)	(0.258)	(0.224)	(0.273)	(0.236)	(0.221)	
Observations	329	46	48	60	47	62	66	

Notes: p-value in column (8) is for a one-way ANOVA on ranks (Kruskal-Wallis) test comparing the six treatment groups in columns (2) to (7). DG refers to the dictator game, in which we gave 20 experimental tokens to participants and asked them how many they would like to keep. Charity rating refers to the rating that subjects gave to the charity that they chose to donate to. We asked subjects to agree to the statement "I like the idea of donating money to [chosen charity]" on a 5-point Likert scale where 1 is "strongly disagree" and 5 is "strongly agree". College major is a categorical variable that summarizes the departmental affiliation of our student subjects.

3.4 Results

Recall that in each of the three rounds of the donation task, subjects can decide to participate in or skip the task. In our discussion of results, we consider each participation in the task as one "donation" (all subjects who choose to participate in the donation task complete it). Participation in the first round of the donation task lets us measure the extensive margin of the donation decision. By summing the number of donations across all three rounds, we can additionally analyze an intensive margin of the decision to donate.

Table 3.4 summarizes those measures and gives an overview of donation behavior across treatments. Panel I presents the fraction of subjects who decide to participate in

each round while panel II sums the number of rounds that subjects decide to participate in the donation task. For subjects in the dual market *CHOOSE* treatment, columns (4) and (5) report whether subjects choose to be paid. In line with our theoretical predictions, donation behavior in the single market *PAID* and the dual market *CHOOSE* treatments is statistically indistinguishable (column 6), both on the extensive margin and the intensive margin.

Table 3.4: Summary Statistics of Behavior in Donation Task
(Fractions and Means, Standard Errors in Parentheses)

	Incentive Treatment			Incentive Choice		p-value
	Not paid (1)	Paid (2)	Choose (3)	Not paid (4)	Paid (5)	H_0: Paid=Choose (6)
I. Fraction of subjects that participated in the task						
			a) PRIVATE treatment			
Round 1	0.609	0.604	0.667	0.083	0.583	0.504
	(0.072)	(0.071)	(0.061)	(0.036)	(0.064)	
Round 2	0.174	0.396	0.467	0.083	0.383	0.463
	(0.056)	(0.071)	(0.065)	(0.036)	(0.063)	
Round 3	0.348	0.313	0.383	0.067	0.317	0.446
	(0.070)	(0.067)	(0.063)	(0.032)	(0.061)	
Observations	46	48	60	60	60	
			b) PUBLIC treatment			
Round 1	0.766	0.806	0.818	0.136	0.682	0.866
	(0.062)	(0.050)	(0.048)	(0.043)	(0.058)	
Round 2	0.383	0.565	0.591	0.136	0.455	0.763
	(0.071)	(0.063)	(0.061)	(0.043)	(0.062)	
Round 3	0.362	0.484	0.530	0.136	0.394	0.601
	(0.070)	(0.064)	(0.062)	(0.043)	(0.061)	
Observations	47	62	66	66	66	
II. Average total number of rounds participated in the task						
			a) PRIVATE treatment			
Sum of all 3 rounds	1.130	1.313	1.517	0.233	1.283	0.290
	(0.129)	(0.142)	(0.135)	(0.072)	(0.132)	
Observations	46	48	60	60	60	
			b) PUBLIC treatment			
Sum of all 3 rounds	1.511	1.855	1.939	0.409	1.530	0.545
	(0.124)	(0.121)	(0.127)	(0.105)	(0.136)	
Observations	47	62	66	66	66	
			c) Aggregating over both visibility treatments			
Sum of all 3 rounds	1.323	1.618	1.738	0.325	1.413	0.348
	(0.092)	(0.095)	(0.094)	(0.066)	(0.096)	
Observations	93	110	126	126	126	

Notes: Total sample size is 329 subjects. Subjects can always choose between participating in the donation task or skipping. P-value in column (6) is for two-sample Wilcoxon rank-sum (Mann-Whitney) test comparing the outcomes for *PAID* treatment in column (2) and the *CHOOSE* treatment in column (3).

In the rest of this section, we pool together observations from *PAID* and *CHOOSE*

treatments to estimate the effects of the availability of incentives on donations behavior. We use this pooled data to provide parametric tests of Hypotheses 1 and 2 on the intensive margin.[11] We then use data from the dual market *CHOOSE* treatment to test Hypothesis 3, again on the intensive margin. We test our three hypotheses on the intensive margin due to better statistical power. Results are qualitatively similar on extensive margin based on the first round of the donation task. In addition to tests of our theoretical hypotheses, we discuss the potential interaction between incentive and visibility effects and analyze heterogeneous treatment effects across genders.

3.4.1 Incentive Effects, Social Image Effects, and Sorting

We test our first two hypotheses in a regression framework. Given the count nature of the outcome variable we use maximum likelihood to estimate the following Poisson regression:

$$Donations_i = \alpha + \beta_1 PAID\&CHOOSE_i + \beta_2 PUBLIC_i + \beta_3 PAID\&CHOOSE_i \times PUBLIC_i + \boldsymbol{X}_i\boldsymbol{\gamma} + \psi_i \tag{3.2}$$

where $Donations$ is the total number of donations by subject i over all three rounds of the donation task, *PAID&CHOOSE* is a dummy for the pooled single market *PAID* treatment and the dual market *CHOOSE* treatment, *PUBLIC* is a dummy for the treatment in which subjects have to reveal their actions to other participants, $\boldsymbol{X}$ is a vector of controls, and ψ is a Poisson-distributed error term. Table 3.5 presents average marginal effect estimates while Appendix Table C.2.1 presents the full set of estimated semi-elasticities.

Our results confirm our first behavioral hypothesis, which says that irrespective of whether actions are visible, the availability of incentives increases donations. We find that compared to the single market *NOT PAID* treatment, the availability of incentives does not induce lower participation in the donation task. This is true irrespective of the visibility of actions. The estimated average marginal effect in our specification with-

[11] Online Appendix C establishes the same results using non-parametric tests.

out any other controls indicates that making incentives available leads to an increase of 0.364 donations over all three rounds (relative to a mean of 1.32 donations in the single market *UNPAID* treatment). The effect size is robust to various sets of controls. Introducing the number of tokens kept in the dictator game as an additional control (Table 3.5, columns 3 to 5) reveals that this measure of altruism is a strong predictor of participation in the donation task.

Result 1 (Incentive Effect). *Irrespective of whether actions are visible, the availability of incentives increases donations.*

Table 3.5: Poisson Regression for Total Donations: Average Marginal Effects (Coefficient Estimates and Standard Errors in Parentheses)

Dependent variable:	# of donations over the three rounds				
	(1)	(2)	(3)	(4)	(5)
	a) Treatments				
PAID&CHOOSE *(Baseline: NOT PAID)*	0.364***	0.360***	0.432***	0.430***	0.456***
	(0.124)	(0.124)	(0.117)	(0.117)	(0.117)
PUBLIC *(Baseline: PRIVATE)*	0.454***	0.462***	0.498***	0.499***	0.494***
	(0.112)	(0.111)	(0.103)	(0.102)	(0.102)
	b) Controls				
Female		0.238**		0.075	0.030
		(0.116)		(0.110)	(0.110)
DG: Tokens kept			-0.099***	-0.097***	-0.090***
			(0.013)	(0.013)	(0.014)
Other Controls	No	No	No	No	Yes
Observations	329	329	329	329	329

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors are clustered at individual level. *NOT PAID* is the base market design treatment. *PRIVATE* is the base visibility treatment. DG refers to the dictator game, in which we gave 20 experimental tokens to subjects and asked them how many they would like to keep. Other controls are age, chosen charity, and individual rating of chosen charity. Note that due to the presentation of average marginal effects, the interaction (which cannot vary independently) is omitted.

We also find support for our second hypothesis of social image effects. Using the same Poisson regression in Equation (4.1), we find that irrespective of the incentive treatment, making actions visible significantly increases the number of donations over all three rounds. The effect of visibility is of similar magnitude to the incentive effect and is similarly robust to various sets of controls.

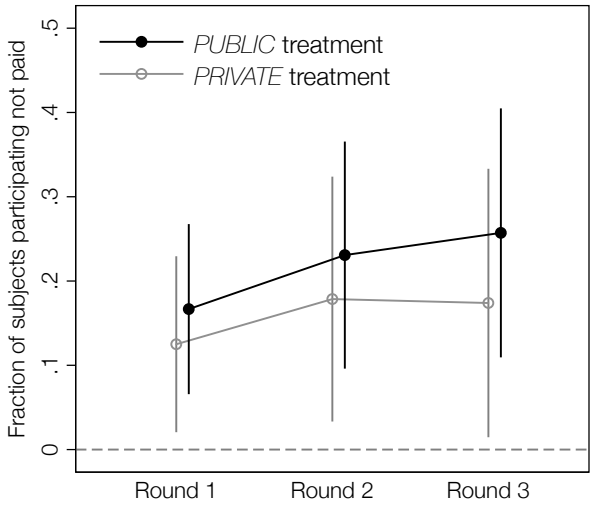
Result 2 (Social Image Effect). *Irrespective of whether compensation is available, making actions visible increases donations.*

We can use our experimental design to assess the potential interaction between incentive and visibility effects. A prominent result in the literature on charitable giving is that incentive effects negatively interact with image effects (Ariely, Bracha, and Meier, 2009). In our framework, in contrast, we do not find a negative interaction between image effects and incentive effects. In the presence of a salient outside option, small incentives to donate do not appear to spoil the image of donors. Appendix Table C.2.1 presents semi-elasticities estimated from Equation (4.1), including for the interaction-term. We estimate a zero interaction effect that is robust across specifications.

Finally, our third behavioral hypothesis states that in a dual market, a positive fraction of donors chooses to be not paid, and that this fraction is larger when actions are observable. We can test this hypothesis by looking at all subjects in the dual market *CHOOSE* treatment.

In each of the three rounds and in each visibility treatment, the fraction of subjects deciding to not be paid for their donation is significantly larger than zero (Figure 3.2). Aggregating over the three rounds, subjects choose to make 0.23 donations without being paid in *PRIVATE* and 0.41 donations without being paid in *PUBLIC* (Table 3.4, panel II, column 4). This confirms the first part of our third hypothesis.

Figure 3.2: Fraction of Participating Subjects Turning Down Incentive in Donation Task, by Round



Notes: Bars indicate 95% confidence intervals. Standard errors clustered at the individual level.

Result 3 (Sorting into unpaid). *In a dual market, a positive share of agents chooses to be not paid.*

In order to analyze sorting into unpaid donations in the dual market *CHOOSE* treatment across visibility conditions, we estimate the following multinomial logit random effect model for the donation decision and the chosen incentive scheme. Each subject i takes decision $d_i \in \{\text{no participation, unpaid participation, paid participation}\}$:

$$d_{i,t} = \alpha + \beta PUBLIC_i + \boldsymbol{X}_i\boldsymbol{\gamma} + v_{i,t} \tag{3.3}$$

where for each subject i and round t, *PUBLIC* is a dummy for the treatment in which subjects have to reveal their actions, $\boldsymbol{X}$ is a vector of controls, and $v_{i,t} = c_i + u_{i,t}$ is the error term of the random effect model. Treatment assignment is permanent, but exogenous. While time invariance of treatment assignment makes the fixed effect model unidentifiable, exogenous treatment assignment meets the random effect assumption and makes this model specification the natural choice.[12]

The multinomial logit random effect model provides estimates for the relative probability of observing not paid rather than paid donations in the *CHOOSE* treatments. In the regression specification without controls, the relative probability increases by 77.3% when actions are visible, and the effect size is fairly stable in specifications with controls (see Table C.2.2). While this confirms qualitatively the pattern from Figure 3.2, this increase is not statistically significant. We are not powered to detect a relative risk ratio that is significantly different from unity at any conventional confidence level.

3.4.2 Heterogenous Social Image Effects Across Genders

We find gender-specific effects in the *PUBLIC* treatment that suggest a differential willingness to engage in costly signaling: Making actions visible increases participation in the donation task significantly among men in the *NOT PAID* and *CHOOSE* treatment. For women, we find the inverse in that the increase is only significant in the *PAID*

[12] Any specification of the regression equation that includes individual characteristics is prone to bias and would require testing of the random effects assumption.

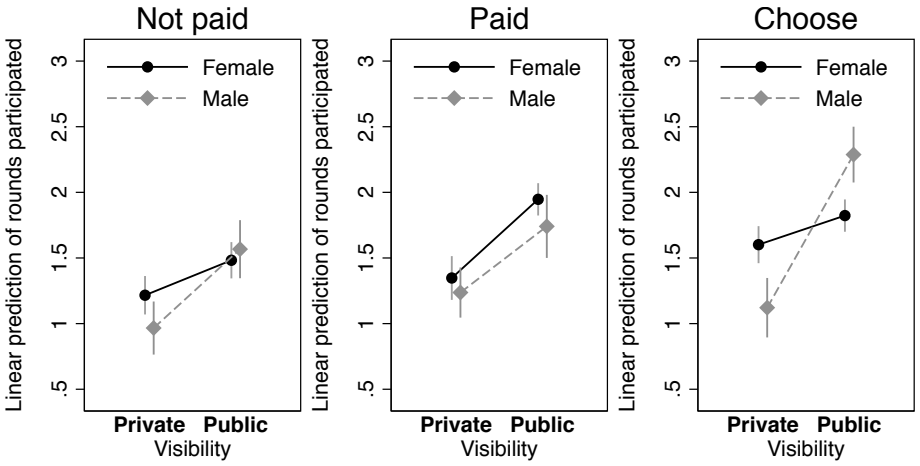
treatment.

Paralleling the analysis above, we use maximum likelihood estimates of a Poisson regression. For each incentive treatment, we separately estimate a model of the form:

$$Donations_i = \alpha + \beta_1 FEMALE + \beta_2 PUBLIC + \beta_3(FEMALE \times PUBLIC) + \beta_4 DG + \psi_i \quad (3.4)$$

where for each subject i, $Donations_i$ is a count variable for number of individual donations over the three rounds of the donation task, and *DG* is the number of tokens kept in the dictator game. Table C.2.3 presents estimates of the semi-elasticities, which reveal that the social image is significantly different across genders only in the dual market *CHOOSE* treatments. Figure 3.3 provides graphical illustration of the interaction effect by plotting the predicted participation in the donation task for each subsample. The heterogeneous effect of public image is particularly salient in the dual market *CHOOSE* treatment.

Figure 3.3: Gender-Specific Effects of Visibility Treatment, by Incentive Treatment (Linear Prediction of Rounds Participated, Based on Regressions in Table C.2.3)



Notes: Bars indicate 95% confidence intervals. Standard errors clustered at the individual level.

We take this as suggestive evidence that men are more willing than women to engage in costly signaling. Recall that in our framework, choosing to participate in the donation task represents a signal that is differentially costly across the three donation

treatments. Choosing to participate without being paid (either in the *NOT PAID* or *CHOOSE* treatments) carries the largest reputational gains, since subjects who engage in the real effort task incur the highest opportunity cost by leaving all value to the charity (i.e. they forego the outside option). In the *PAID* treatment, subjects can signal their altruism at a lower opportunity cost (i.e. they forego the outside option minus the individual compensation).

3.5 Discussion and Conclusion

Motivated by the market for blood donations in Germany, where different incentives for altruism coexist and donors can effectively turn down monetary incentives to donate, we set out to study a "dual market" for the collection of charitable donations. While incentives for prosocial behavior have mostly been studied in isolation and contrast to the absence of incentives, we explicitly allow agents to turn down a compensation for their donation.

In the case of blood donations in Germany, different blood collectors offer different incentives and prospective donors can choose where to donate. Donations at the Red Cross are always unpaid, while donations at hospitals or commercial blood banks are compensated with 20 to 30 euro. Everyone who lives in one of the 50 largest communities in Germany can reach an unpaid donation point of the Red Cross within 30 minutes time driving or on public transport. This compares to about 62% of the population who can reach a paid donation point within 30 minutes time using the same means of transport (see in the online appendix Table E2 for details and and Figure E1 for the spatial distribution of blood collection centers). In Meyer and Tripodi, 2018 we survey knowledge of various institutions to donate blood in the city of Bonn and find awareness for paid and unpaid options to be similar (see in the online appendix Table E3).[13] While donors appear to be able to choose whether or not they want to be paid,

[13]Meyer and Tripodi, 2018 interview about 1,000 randomly sampled customers of the municipal service center in Bonn, a mid-sized city in the west of Germany. Although the data is not representative for Germany, we take awareness of both paid and unpaid collection centers, for a rich set of demographic groups in an urban area, as confirmation that the choice between incentives for donating blood is indeed salient for a non-negligible share of the population.

unpaid donations still represent more than 70% of all donations in Germany (Paul-Ehrlich-Institut, 2018). Incidentally, the German market also has the highest per capita rate of donations among all 172 countries that report to the WHO and comparatively low wholesale prices for human blood.[14]

We study such a dual market in a stylized environment. The results from our laboratory experiment support our three behavioral hypotheses. We confirm our first hypothesis, which predicts that introducing a compensation for a donation should increase giving. In the dual market, the availability of extrinsic incentives does not crowd out intrinsic motivations of donors. In fact, giving significantly increases compared to the market design in which donations are not paid. These findings stand in contrast with the influential work of Titmuss, 1971, who argued that paid blood donations could crowd out the intrinsic motivation to donate and lead to a net drop in donations.

For a simple illustration of the effect size, we can use the average marginal effects from the Poisson regression of the number of individual donations over the three rounds on treatment indicators, a gender dummy, and the number of tokens kept in the dictator game (Table C.2.4, column 5). Holding everything else constant, the predicted number of donations in a dual market is 0.473 standard deviations larger than in the single market where donations are not paid. This is equivalent to the estimated effect of moving from the 20th percentile to the 60th percentile in the distribution of "generosity" of subjects as measured by the dictator game, again holding everything else constant.

Offering a compensation and letting agents turn down the compensation lets the collection system leverage the heterogeneity in individual preferences. This enables efficiency gains in the collection similar to those deriving from self-selection in second-

[14]Germany has the highest number of donations at 57.3 per 1,000 people, compared to 49.2 in Sweden and 43.7 in the United States. The cost of one blood unit on the German wholesale market is among the lowest in the world at about $110, compared to $190 in Sweden and Switzerland (Trimborn, 2009) and about $211 in the United States (Toner et al., 2012). We calculate per capita donations based on the total number of whole blood donations collected in the years 2011 to 2013 (World Health Organization, 2017). We use the latest year available for all countries that report to the WHO. Population data comes from the World Bank World Development Indicator database. Online appendix E provides more details on the German market for whole blood donations.

degree price discrimination. Our sorting hypothesis states that in a dual market, a positive fraction of donors chooses to be not paid and that this fraction is bigger when actions are taken in public. We find that when given the option to turn down the compensation, a significant fraction of donors choose to do so, though we find only weak evidence that donors turn down incentives more in public than in private. This result complements the findings of Lacetera, Macis, and Slonim, 2014, who conduct a field experiment in which the American Red Cross offers gift cards as incentive to donate blood. They report that after donating, virtually none (2%) of the offered cards were turned down. In their setting, the ability to turn down incentives is not salient to prospective donors in their decision to come to the donation drive. Moreover, there is no clear signaling motive for turning down the gift card. In our setting, the two incentive schemes carry different utility in terms of private benefit and signaling value. With this choice between the two different incentives schemes, our dual market should be more effective at leveraging heterogeneity in individual preferences.

Even though we cannot provide strong evidence that sorting operates through social image concerns, we do find robust support of our second hypothesis, which states that visibility of actions increases donations irrespective of the type of available incentives. We can again use the average marginal effects from Poisson regression (Table C.2.4, column 4) to illustrate the effect size of social image. Making actions observable while holding everything else constant increases the predicted number of donations by 0.493 standard deviations. This is slightly larger than the estimated effect of moving from the 20th percentile to the 60th percentile in the distribution of "generosity" of subjects as measured by the dictator game, again holding everything else constant.

The single market *PAID* and *NOT PAID* treatments allow us to compare our findings to the existing literature. In contrast to previous work, we do not find that social image effects attenuate incentive effects (Ariely, Bracha, and Meier, 2009; Carpenter and Myers, 2010). Individuals in our experiment have an outside option that is larger than the monetary incentives to donate, so that *homo economicus* would never choose to

donate. Both our work and Ariely, Bracha, and Meier, 2009 are based on the theoretical framework of Benabou and Tirole, 2006. Our findings suggest that in this framework, a salient outside option makes incentivized donations more likely to signal altruism and less likely to signal greed. This attenuates the image-spoiling effects of incentives that can bring about a negative interaction between incentive and image effects.

Our findings also suggest a gender-specific willingness to engage in costly signaling that could be interpreted as consistent with gender-specific aversion to standing out (Jones and Linardi, 2014) as well as with costly signaling theory in evolutionary biology (Gintis, Smith, and Bowles, 2001; Smith and Bird, 2000) and strategic signalling of generosity among men (Eagly and Crowley, 1986; Iredale, Van Vugt, and Dunbar, 2008; Barclay, 2010; Boehm and Regner, 2013).

Our findings have implications for the design of mechanisms for the collection of charitable donations. Applied to the collection of whole blood donations, our results could inform the design and regulation of systems that use monetary incentives. Because voluntary provision of blood donations is often insufficient (Whitaker et al., 2016), demand for blood is likely increasing in the future (Greinacher et al., 2011), and modern screening technologies appear sufficiently safe to counter adverse selection (Offergeld et al., 2005), several countries are now re-evaluating partial reliance on incentivized or paid donations (Lacetera, Macis, and Slonim, 2013). Even small efficiency gains in these collection systems can imply economically meaningful savings for public health budgets. In the Unites States alone, about 13.6 million blood units are collected every year at a total value of more than US$ 3 billion.[15] Our results suggest that having different institutions provide distinct incentive schemes can improve the efficiency of the market compared to the case of all institutions offering the same incentives. In such a market, collectors may be able to increase donations by making image concerns more salient. In the case of Germany, the institution that offers unremunerated donations and has most to gain from making donations visible – the Red Cross – in fact largely relies on highly visible mobile drives for its collection.

[15]Back-of-the-envelope calculation based on 2007 US data from Toner et al., 2011.

Our results point to various avenues for future research. First, it would be good to better understand the mechanisms through which sorting into unpaid donations operates both in the German blood market and in general. While our theoretical framework suggests that social image effects should play a key role, our experimental data provides only weak evidence to support this hypothesis. Second, our setting does not appear to suffer from the negative interaction of social image effects and incentive effects that has been found in the previous literature. Empirical studies to determine if and when incentives spoil image utility constitute fruitful avenue for future research. Third, we cannot rule out that specific features of our experimental task undermine the external validity of our findings. While we used a task that is popular in the literature because it is not inherently meaningful and lends itself to a test of subject motivation, there is scope for future work in less stylized settings. Finally, we hope this work stimulates theoretical efforts on the characterization of competitive aspects of dual markets that would allow us to better understand the endogenous formation and social welfare implications of such institutional arrangements–important matters from which we largely abstract in this paper.

Appendices

Appendix A

Appendix to Chapter 1

A.1 British Parliamentary debating

Debates can take place in various formats. The most popular format, that features in the most prestigious tournaments (e.g. the World University Debating Championship), is the British Parliamentary (BP). For such format, debaters take part in debates in teams and each team is composed of two debaters. A debate is characterized by a motion, four teams of debaters, and a panel of experienced judges. Debates begin with the announcement of the motion that two teams, on the proposition (also called Government) side of the House have, to defend and two teams, of the opposition side of the House, have to contrast. BP debating exclusively feature *impromptu* debates, in which motions are revealed only 15 minutes ahead of debates and teams are randomly assigned to argue either in favor or against the given motion. Finally, while the order of teams speaking in each debate is also random, it is each team's choice to determine which team member speaks first. All speakers are given 7 minutes to present their arguments following a precise structure that we illustrate in Table A.1.

Table A.1: Debaters' Responsibilities by Role

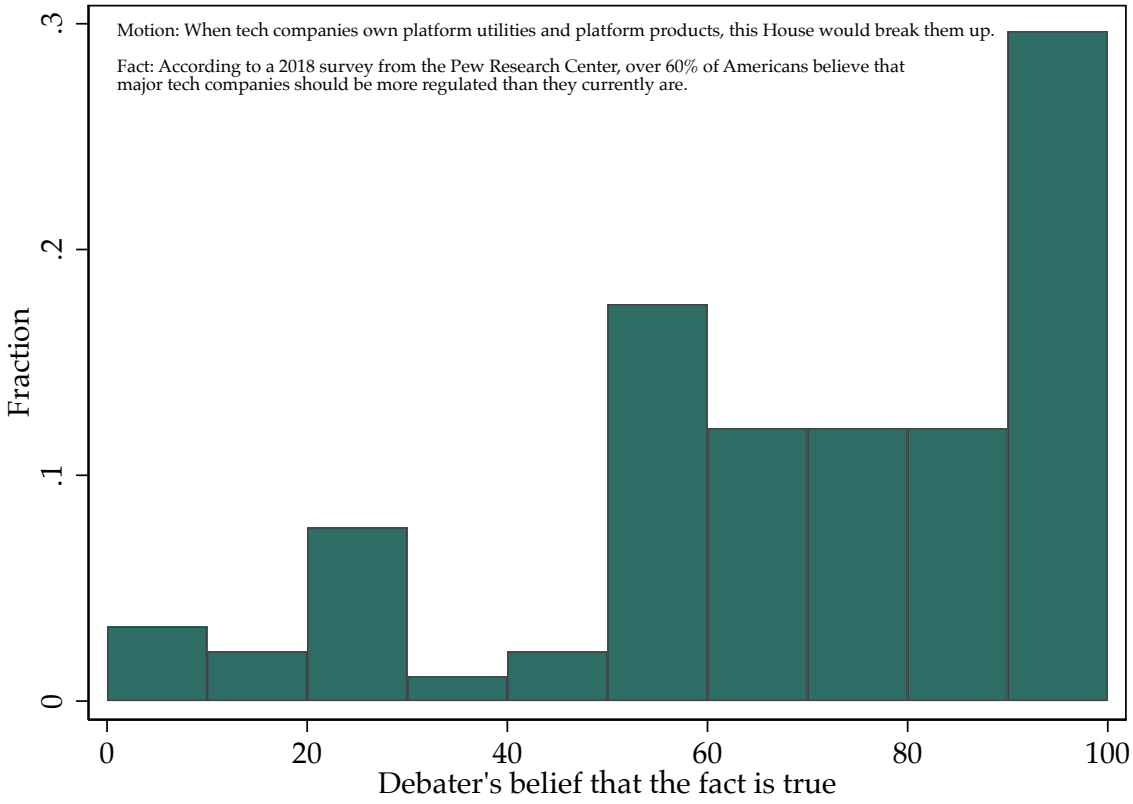
Team	Speaking role	Speaking order	Team	Speaking role	Speaking order
Opening Government (OG)	Prime Minister (PM)	First	Opening Opposition (OO)	Leader of the Opposition (LO)	Second
	• Defines and interprets the motion • Develops the case for the proposition			• Accepts definition of the motion • Refutes the case of OG • Constructs arguments against PM's interpretation of the motion	
	Deputy Prime Minister (DPM)	Third		Deputy Leader of the Opposition (DLM)	Fourth
	• Refutes the case of OO • Rebuilds the case of OG • May add new arguments to the case of the PM			• Continues refuting the case of OG • Rebuilds the case of OO • May add new arguments to the case of the LO	
Closing Government (CG)	Member of the Government (MG)	Fifth	Closing Opposition (CO)	Member of the Opposition (MO)	Sixth
	• Defends the general direction and case of OG • Continues refutation of OO • Develops a new argument that is different from but consistent with the case of OG			• Defends the general direction taken by OO • Continues general refutation of OG's case • Provides more specific refutation of CG's case • Provides new opposition arguments	
	Government Whip (GW)	Seventh		Opposition Whip (OW)	Eighth
	• Summarizes the entire debate from the point of view of the proposition, defending the general view point of both OG and CG with a special eye toward the case of CG • Does not provide new arguments			• Summarizes the entire debate from the point of view of the proposition, defending the general view point of both OO and CO with a special eye toward the case of CO • Does not provide new arguments	

A.2 Example Motion, Factual Belief Questions, and Attitudes Elicitation

For every motion, we devise four factual statements and two charitable donations tailored to the motion.

All facts are based on exact statistics from high quality research/reports/surveys. Instead of exact statistics, we report to subjects broad intervals, including values either above or below a given threshold, within which the exact statistic may or may not fall into. This allows us to formulate binary statements for which we ask debaters to predict whether the statement is true or false. Factual statements are devised in a way that truths that appear *convenient* on one side of the debate are instead *inconvenient* on the opposite side. Figure A.1 presents one of the four factual statements devised for a motion on breaking up big tech companies, and provides the distribution of elicited beliefs. This factual statement was devised expecting that it would be convenient for a speaker arguing *in favor* of the motion if the statement were true, and convenient for a speaker arguing *against* the motion if it were false. For both tournaments we collect 36 factual questions related to the motion. About half of these factual statements are favorable to the proposition (opposition) if true.

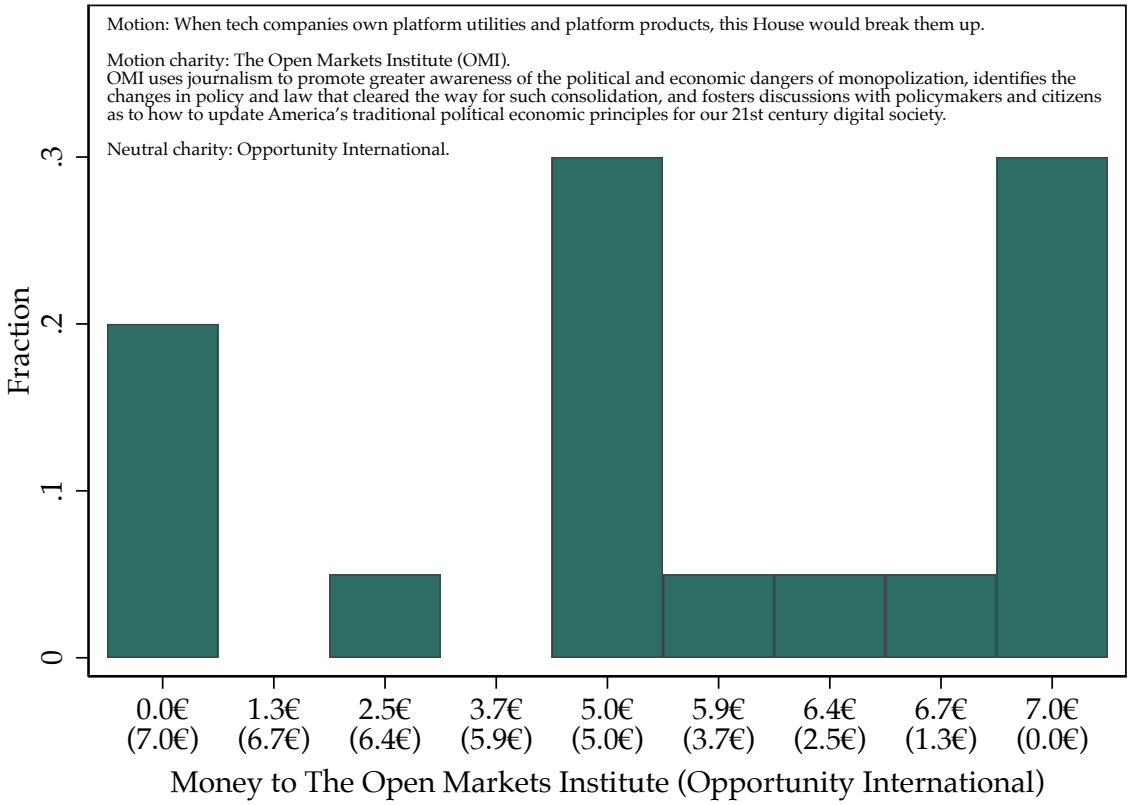
Figure A.1: Example Distribution of Reported Beliefs on a Factual Statement



All charities related to the motion are selected NGOs such that individuals on one side of the debate, who are truly convinced of the factual and moral merits of their persuasion goal, would tend to favor the charitable cause related to the motion. Figure A.2 presents one of the two motion charities devised for the motion on breaking up big tech companies, and provides the distribution of elicited monetary allocations. In this case, we expected individuals who would genuinely argue the proposition (opposition) side of the debate to display a relative preference for the motion charity (neutral charity). The choice of charities is restricted to NGOs that have no known (or alleged) relationship with terrorist organizations.[1]

[1]Non-trivial restriction given that two of the nine motions were explicitly related to terrorism.

Figure A.2: Example Distribution of Chosen Monetary Allocations Between a Motion-Specific Charity and a Neutral Charity



A.3 Belief and Attitude Convergence

In this section we present estimation of two quantitative measures of cultural polarization. First, we present estimates of an axiomatized index of polarization for continuous distributions (Duclos, Esteban, and Ray, 2004). Second, we present estimates of an index of cultural distance, borrowed by economists from population genetics, that incorporates socio-demographic information to assess distance along a particular dimension across cultural labels (Desmet, Ortuño-Ortín, and Wacziarg, 2017).

The first measure of polarization, reflects an identification-alienation framework of conflict, in which polarization and conflict are intimately related, and conflict in society stems from alienation across individuals and proximity within groups of individuals that are alienated from the rest of society. This measure ignores cultural labels, but rather incorporates identities as modal observations of the variable of interest y.

$$P_\alpha(y) = \int \int f(y)^{1+\alpha} f(y')|y - y'|dy\ dy'$$

for $\alpha \in [0.25, 1]$ polarization sensitivity parameter.

The second measure of polarization Φ_{ST}, incorporates cultural labels to capture the extent to which, along the outcome of interest y, individuals within a certain group are similar to one other relative to overall similarity in the population. Such index is obtained as

$$\Phi_{ST}(y) = \frac{P_0(y) - \sum_{g \in G} w_g P_0(y)_g}{P_0(y)}$$

where $P_0(y)$ is the polarization index estimated at $\alpha = 0$, g denotes a cultural label in the set of cultural labels G, w_g is the share of individuals in the population with cultural label g, and $P_0(y)_g$ is the polarization index computed for the distribution of y among individuals in group g at $\alpha = 0$.

Table A.2: Cultural Distance and Polarization, by Question and Survey

Motion	Φ_{ST}				P^2			
	Base (B)	Pre (P)	Post (B)	Post (P)	Base (B)	Pre (P)	Post (B)	Post (P)
1	0.028	0.010	0.007	0.018	0.288	0.330	0.285	0.279
	0.011	0.008	0.005	0.015	0.315	0.300	0.300	0.313
2	0.014	0.022	0.021	0.032	0.284	0.309	0.292	0.310
	0.024	0.070	0.019	0.043	0.326	0.323	0.311	0.310
3	0.006	0.080	0.021	0.078	0.285	0.280	0.279	0.298
	0.006	0.035	0.008	0.019	0.297	0.316	0.294	0.299
4	0.005	0.018	0.012	0.005	0.295	0.281	0.291	0.272
	0.014	0.010	0.007	0.010	0.280	0.287	0.300	0.308
5	0.010	0.007	0.002	0.022	0.304	0.326	0.291	0.277
	0.004	0.010	0.039	0.019	0.309	0.301	0.275	0.286
6	0.023	0.050	0.016	0.016	0.322	0.300	0.288	0.289
	0.015	0.011	0.108	0.038	0.309	0.296	0.312	0.293
7	0.006	0.069	0.009	0.015	0.303	0.283	0.272	0.280
	0.025	0.033	0.035	0.052	0.315	0.306	0.292	0.282
8	0.015	0.036	0.045	0.061	0.286	0.299	0.311	0.300
	0.022	0.046	0.008	0.019	0.312	0.335	0.298	0.278
9	0.004	0.024	0.017	0.026	0.284	0.288	0.297	0.322
	0.008	0.075	0.030	0.011	0.305	0.294	0.292	0.281
Average	**0.013**	**0.034**	**0.023**	**0.028**	**0.301**	**0.293**	**0.303**	**0.293**
95% CIs	[0.010 – 0.017]	[0.023 – 0.046]	[0.012 – 0.034]	[0.019 – 0.037]	[0.294 – 0.308]	[0.288 – 0.299]	[0.295 – 0.311]	[0.286 – 0.300]

Notes: Confidence intervals around the average of each index across questions are obtained from 500 simulated bootstrap samples of the indices underlying the average. Base (B) [Post (B)] refers to indeces computed on answers collected from questions that are only asked at baseline [postdebate]. Pre (P) [Post (P)] refers to indeces computed on answers collected from questions that are only asked at predebate [postdebate].

Table A.2 shows relatively little cultural distance across proposition and opposition speakers, and moderate polarization along elicited beliefs.

The bottom raw of the table aggregates the indices computed at the question-survey level to make inference about how debates affect these measures. We find that on average polarization increases from baseline to postdebate, and remains constant from predebate to postdebate. This suggests that debates can increase polarization because of self-persuasion, and the exchange of views taking place during debates may be ineffective at driving a social consensus.

Cultural distance increases from baseline to postdebate, and decreases (by a somewhat smaller extent) from predebate to postdebate. These patterns confirm that self-persuasion drives beliefs apart between proposition and opposition speakers, and

show that the exchange of views can play some role in reducing divergence.[2]

Table A.3 shows that the debate helps speakers form beliefs that are closer to the truth ((1) and (2)). Columns (3) to (11) provide the simplest possible tests of beliefs and attitude convergence that were included in the pre-analysis plan. The results are largely consistent with the main analysis presented in Section 1.3: at the individual level, (i) distance from median belief is larger at postdebate than it is at baseline, (ii) distance from median belief is not statistically different between postdebate and predebate, and (iii) the same is for distance from median chosen charity allocation bundle. Columns (7), (8), and (11) indicate that even if we restrict the analysis to the half of the sample of subjects whose beliefs at baseline are aligned to the randomly assigned persuasion goal we observe similar qualitative patterns as for the full sample. This analysis is however only very suggestive as we are clearly under-powered to detect significant convergence/divergence in this sub-sample.

Table A.3: Fixed Effect Regression for Convergence in Beliefs and Attitudes

	Distance from Truth		Distance from Median								
			Beliefs						Charity allocation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Baseline survey	4.152**		-1.429*	-1.214			-1.708				
relative to Postdebate	(1.654)		(0.836)	(1.204)			(1.262)				
Predebate survey		1.998			0.953	0.810		1.907	0.002	0.001	-0.002
relative to Postdebate		(1.478)			(0.813)	(1.090)		(1.258)	(0.055)	(0.080)	(0.071)
Baseline survey × Heated debate				-0.402						0.000	
				(1.656)							
Predebate survey × Heated debate						0.267				0.002	
						(1.512)				(0.105)	
Heated debate				0.837		3.367***				0.027	
				(1.284)		(1.276)				(0.110)	
Observations	1753	1769	1753	1753	1769	1769	856	855	1766	1766	854

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Notes: Heated debate is a binary variable indicating, for each round of debate, the debates in which the average subjective heat score of speakers in a debate room is above the median. Standard errors in parentheses are clustered at the team level.

A recurrent finding in social psychology and political science is that the exchange

[2]Unfortunately, by design, we can only directly compare the estimates of these indices from baseline to postdebate and from predebate to postdebate, as the underlying factual statements on which beliefs are elicited differ for different debaters across these two sets of surveys.

of views can either polarize or unite individuals depending on the level of conflict that surrounds the conversation (see e.g. Mutz, 2007, and references therein). Hence, we interact a measure of conflict in a debate, based on how heated enumerators score single debaters in a debate room to be, with the timing of the outcome elicitation.[3] We would have expected more heated debates to possibly increase polarization and less heated debates to decrease it, but we do not find support for such interaction.

[3]If we instead use for this analysis an objective measure of conflict in a debate, given by the number of times that speakers in a debate are challenged by the opposing teams, we obtain qualitatively similar results.

A.4 Additional Figures and Tables

Table A.4: Debater Characteristics by Tournament

	Full sample	by tournament		
		Munich	Rotterdam	p-value
Female	0.351	0.427	0.261	0.017
	(0.035)	(0.049)	(0.047)	
Age	21.715	21.573	21.878	0.196
	(0.205)	(0.302)	(0.274)	
Time in debating	2.326	2.340	2.311	0.809
	(0.072)	(0.099)	(0.106)	
Past achievements	3.218	2.078	4.522	0.192
	(0.763)	(1.199)	(0.876)	
Local nationality	0.245	0.250	0.239	0.860
	(0.031)	(0.043)	(0.045)	
Left to right political ideology scale	3.372	3.294	3.461	0.734
	(0.134)	(0.173)	(0.208)	
Observations	196	104	92	196

Note: The last column reports the p-value from a one-way ANOVA on ranks (Kruskal-Wallis) test comparing the two tournaments.

Table A.5: Debaters' Baseline Beliefs and Characteristics, by Tournament and Side of the Motion

	Munich				Rotterdam			
	Full sample	Opposition	Proposition	p-value	Full sample	Opposition	Proposition	p-value
				(a) By motion				
Baseline belief motion 1	44.369	45.596	43.118	0.764	52.322	52.022	52.636	0.881
	(3.084)	(4.303)	(4.456)		(3.474)	(5.212)	(4.623)	
Baseline belief motion 2	39.794	36.314	43.275	0.193	51.378	46.854	56.548	0.131
	(3.131)	(4.652)	(4.181)		(3.084)	(4.537)	(4.008)	
Baseline belief motion 3	65.000	64.451	65.549	0.965	39.483	40.907	38.152	0.578
	(2.622)	(3.837)	(3.609)		(3.255)	(4.498)	(4.729)	
Baseline belief motion 4	52.363	51.667	53.059	0.820	56.989	58.444	55.500	0.684
	(2.818)	(3.996)	(4.010)		(3.173)	(4.525)	(4.489)	
Baseline belief motion 5	71.588	72.608	70.569	0.968				
	(2.645)	(3.403)	(4.079)					
Observations	104	52	52		96	48	48	
				(b) All motions				
Female	0.427	0.438	0.416	0.620	0.262	0.258	0.266	0.874
	(0.022)	(0.031)	(0.031)		(0.024)	(0.033)	(0.034)	
Age	21.573	21.519	21.626	0.948	21.877	21.847	21.909	0.703
	(0.134)	(0.183)	(0.197)		(0.137)	(0.194)	(0.193)	
Time in debating	2.340	2.341	2.339	0.981	2.315	2.279	2.352	0.464
	(0.044)	(0.062)	(0.063)		(0.053)	(0.074)	(0.074)	
Achievements	3.069	3.196	2.941	0.583	4.529	4.284	4.784	0.766
	(0.304)	(0.457)	(0.402)		(0.437)	(0.583)	(0.656)	
Local nationality	0.250	0.238	0.263	0.527	0.237	0.246	0.228	0.682
	(0.019)	(0.026)	(0.027)		(0.022)	(0.032)	(0.031)	
Political scale	3.294	3.271	3.318	0.843	3.462	3.497	3.425	0.612
	(0.077)	(0.108)	(0.110)		(0.104)	(0.143)	(0.151)	
Observations	519	259	260		367	175	192	

Note: P-value is from a one-way ANOVA on ranks (Kruskal-Wallis) test comparing the two groups. Each observation is a debater at each round of the tournament. For panel (a) we have a total of 104 observations for each Factual Beliefs relating to the motions of each round. For panel (b), where the outcomes are not round specific while treatment assignment is, the number of observations equals the number of debaters in each position across all rounds of the tournament.

Table A.6: Debaters' Baseline Characteristics, by Tournament

	Munich				Rotterdam			
	Full sample	Group 1	Group 2	p-value	Full sample	Group 1	Group 2	p-value
Female	0.427	0.451	0.404	0.630	0.261	0.349	0.178	0.069
	(0.049)	(0.070)	(0.069)		(0.047)	(0.074)	(0.058)	
Age	21.573	21.667	21.481	0.519	21.878	22.233	21.553	0.282
	(0.302)	(0.422)	(0.435)		(0.274)	(0.417)	(0.357)	
Time in debating	2.340	2.314	2.365	0.732	2.311	2.302	2.319	0.953
	(0.099)	(0.144)	(0.137)		(0.106)	(0.158)	(0.143)	
Achievements	3.069	2.255	3.882	0.223	4.522	4.488	4.553	0.880
	(0.682)	(0.557)	(1.243)		(0.876)	(1.133)	(1.331)	
Local nationality	0.250	0.269	0.231	0.652	0.239	0.227	0.250	0.800
	(0.043)	(0.062)	(0.059)		(0.045)	(0.064)	(0.063)	
Political scale	3.294	3.627	2.961	0.108	3.461	3.738	3.213	0.227
	(0.173)	(0.264)	(0.215)		(0.208)	(0.293)	(0.293)	
Observations	104	52	52		92	44	48	

Note: The two partitions of teams (Group 1 and Group 2) answer the same set of question, but answer sets of factual beliefs and attitude elicitations in different orders across surveys. P-value is from a one-way ANOVA on ranks (Kruskal-Wallis) test comparing the two groups.

Table A.7: Ordered Logit Regressions for Effect of Persuasion Goals on the Allocation of Charitable Donations

	Donation bundle favorable to proposition charity		
	(1)	(2)	(3)
Speaker in proposition	0.271**	0.274**	0.282**
	(0.120)	(0.127)	(0.131)
Socio-demographic and experience controls		✓	
Debater fixed effects			✓
Round fixed effects	✓	✓	✓
Observations	883	850	883

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level for the random effects estimates (columns (1) to (2)), and at the individual level for the fixed effects estimates (column (3)). Fixed effects estimates are obtained from the Baetschmann, Staub, and Winkelmann, 2015 estimator to overcome notorious under-identification problem of ordered logit models with fixed effects Chamberlain, 1980. Socio-demographic controls include age, gender, and an indicator for whether the speaker's nationality is from the country that hosts the competition. Experience controls include the reported number of international tournaments in which the speaker has made it to semi-finals, and a categorical variable capturing the number of years the speaker has been actively debating.

Table A.8: Panel Regressions for Effects of Persuasion Goals, by Gender

	Factual Beliefs		Attitudes		Confidence	
	Female	Male	Female	Male	Female	Male
Debater in proposition	0.255**	0.171**	0.059	0.464**	3.108	4.525**
	(0.099)	(0.075)	(0.229)	(0.180)	(2.745)	(1.852)
Socio-demographic and experience controls	✓	✓	✓	✓	✓	✓
Round fixed effects	✓	✓	✓	✓	✓	✓
Observations	307	544	306	544	307	543

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors in parentheses are clustered at the team level. Socio-demographic controls include age, gender, and an indicator for whether the debater's nationality is from the country that hosts the competition. Experience controls include the number of years the debater has been actively debating.

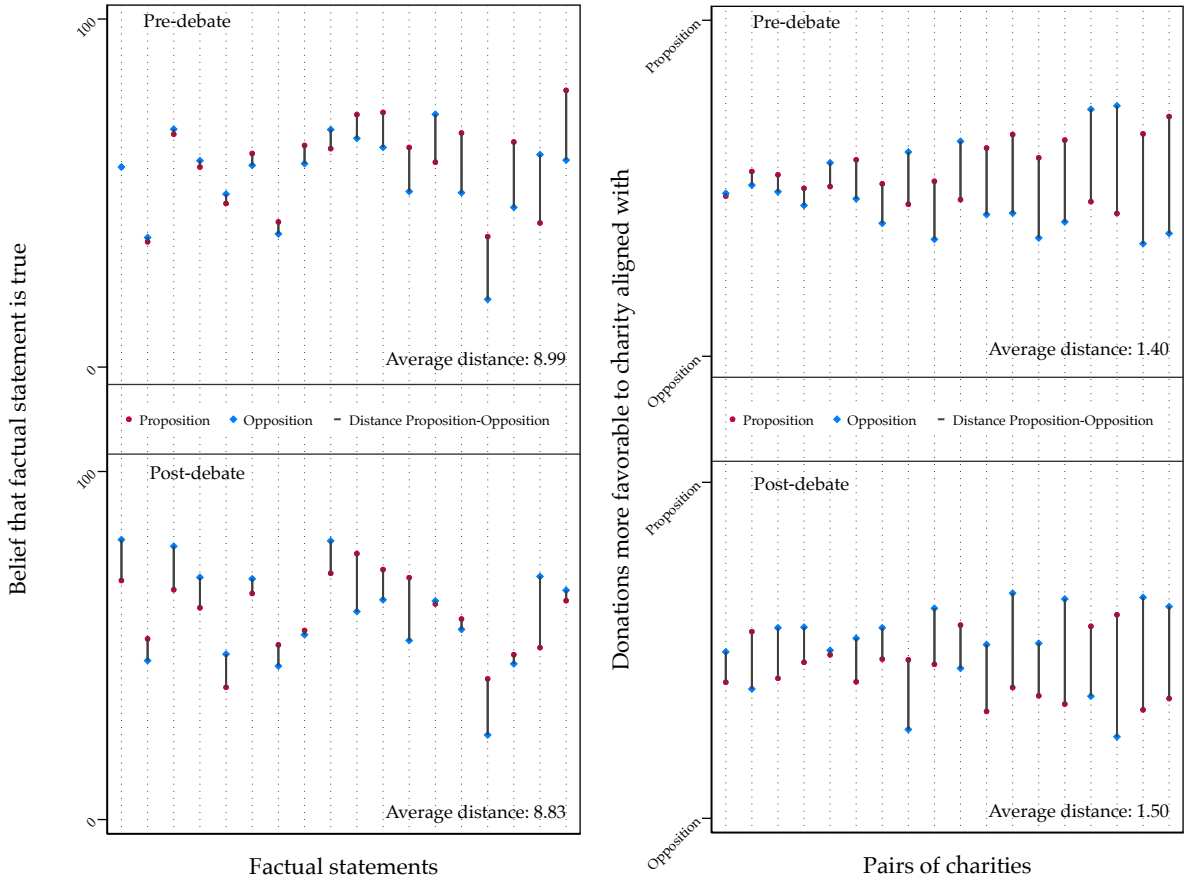
Table A.9: Panel Regressions for Alignment by Position Assigned and by Winning Side

	Belief alignment with Proposition				Attitude alignment with Proposition			
	Predebate		Postdebate		Predebate		Postdebate	
Debater in proposition	0.210***		0.085		0.299**		0.324**	
	(0.062)		(0.081)		(0.139)		(0.155)	
Debate won by proposition team		-0.102		0.202***		0.278*		0.207
		(0.079)		(0.072)		(0.153)		(0.179)
Debater fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Round fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Observations	851	851	850	849	843	843	842	841

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

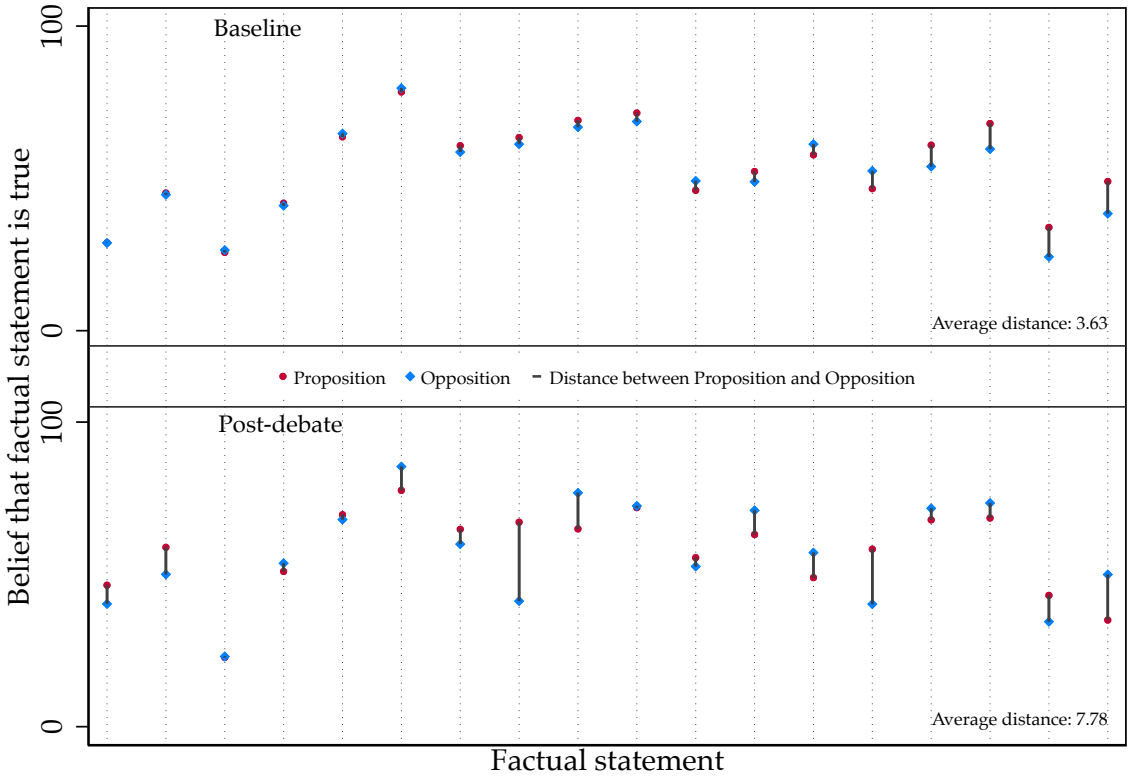
Notes: Standard errors in parentheses are clustered at the team level.

Figure A.3: Distance in Beliefs and Attitudes, Pre- and Post- Debate



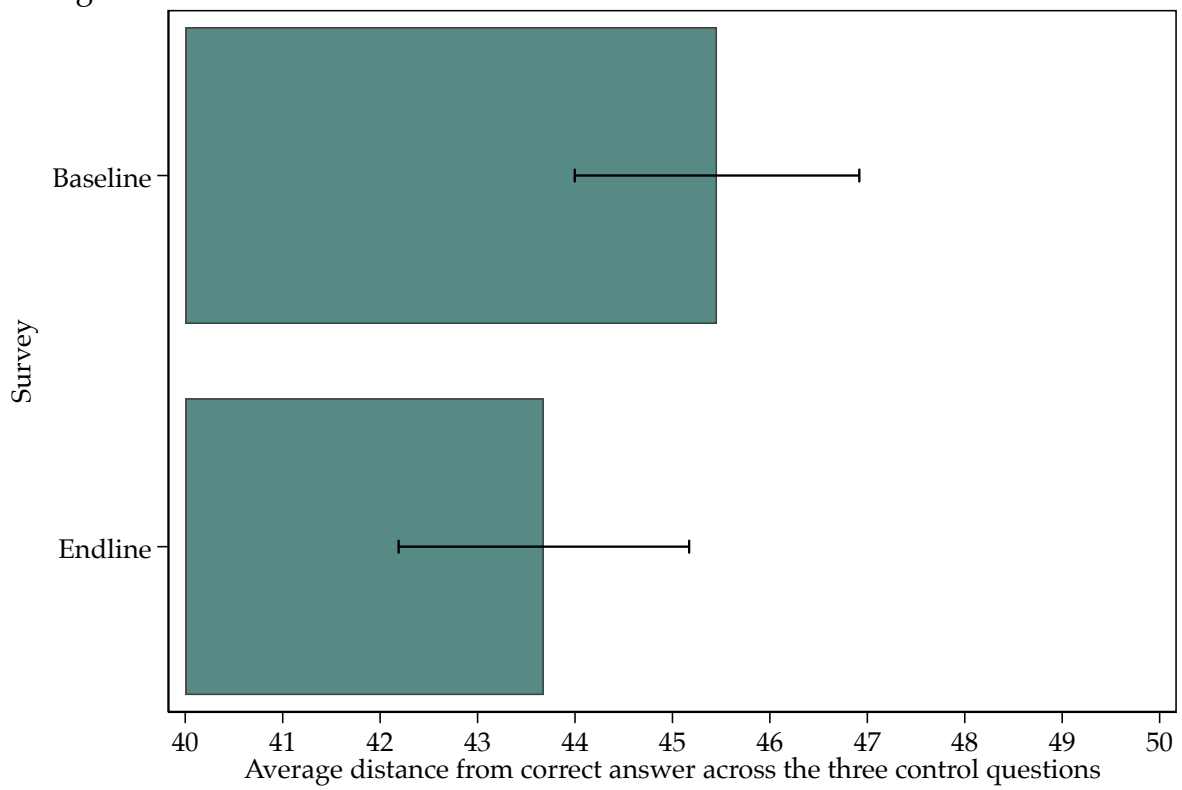
Note: Each vertical dotted line represents either a factual statement over which beliefs are elicited at predebate (top left panel) and postdebate (bottom left panel), or a pair of charities between which debaters allocate monetary endowments at predebate (top right panel) and postdebate (bottom right panel). In the left (right) panel, colored markers represent average report (chosen monetary allocation bundle) among speakers on each side of the debate. Black segments between each pair of colored markers represent the distance in the average position of speakers on the two sides of the debate. For each panel, for readability, factual statements and pairs of charities are sorted by distance between average proposition and opposition outcomes at the predebate stage. The four sets of outcomes are summarized in the bottom right corner by the average distance between the average positions of proposition and opposition.

Figure A.4: Distance in Beliefs, at Baseline and Post- Debate



Note: Each vertical dotted line represents a factual statement over which beliefs are elicited at baseline (top panel) and postdebate (bottom panel). Colored markers represent average report among speakers on each side of the debate. For readability, factual statements are sorted by distance between average proposition and opposition outcomes at the baseline stage. The two sets of outcomes are summarized in the bottom right corner by the average distance between the average positions of proposition and Opposition.

Figure A.5: Evidence on Learning of Correct Answers to Belief Elicitation Questions Through the Entire Tournament



Note: Mean distances of reported beliefs from correct answers are averaged at the individual level for the three control questions in each survey. This figure reports the survey average of such individual-survey level metrics and the corresponding error bars.

A.5 Predictors of Persuasiveness

Table A.10: Panel Regressions fo Correlation Between Persuasiveness and Alignment with the Motion (Standard Errors in Parentheses)

	Broad persuasiveness score				Quality of argumentation score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline belief aligned (binary outcome)	0.009	0.015			0.114	0.092		
	(0.075)	(0.076)			(0.216)	(0.218)		
Baseline belief alignment (continuous outcome)			-0.009	-0.011			0.129	0.109
			(0.034)	(0.036)			(0.110)	(0.110)
Debater fixed effects	✓		✓		✓		✓	
Socio-demographic and experience controls		✓		✓		✓		✓
Round fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Observations	869	848	869	848	869	848	869	848

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Heteroskedasticity robust standard errors in parentheses.

Table A.11: Pair-wise Correlation Between Persuasion Outcomes and Potential Predictors

	Broad persuasiveness	Quality of arguments
	(1)	(2)
(a) Pearson's correlation		
Achievements	0.475***	0.528***
	(0.000)	(0.000)
Factual knowledge at baseline	0.118	0.126*
	(0.102)	(0.080)
Predebate share of strong arguments for the other side of the debate	0.037	0.087
	(0.604)	(0.229)
Predebate share of arguments for the other side of the debate	0.017	0.042
	(0.814)	(0.564)
(b) Spearman's rank correlation		
Time in debating	0.549***	0.479***
	(0.000)	(0.000)
Observations	196	196

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: P-value for statistical significance in parentheses. All variables for this analysis are averaged across all rounds of debate. Broad persuasiveness of a debater is evaluated by each judge on the panel independently; for this analysis we use panel averages of broad persuasiveness. Factual knowledge at baseline captures, how close debaters' beliefs on the 5 motion related factual statements presented at baseline are to the truth. Predebate belief (attitude) alignment captures how close debaters' beliefs are to the response aligned with their persuasion goal.

A.6 Heat of Debates

Table A.12 summarizes our two measures of heat in a debate. The first is an objective proxy obtained by counting how many times a speaker is challenged by non-speaking debaters in the room. The second is a subjective heat score that the enumerator attributes to each speaker in the room. The average of these two individual outcomes at the round level are informative of how much heat each motion generates, and visual inspection of the table already indicated a positive correlation of these two outcomes.

Table A.12: Average Heat Score (Standard Errors in Parentheses)

Motion	Number of POIs	Subjective heat
	(1)	(2)
This House believes that governments should stop funding scientific programs that have	4.165	2.680
no immediate benefit for humankind (such as space travel and exploration, human cloning).	(0.300)	(0.123)
This House believes that Western States should permanently revoke the citizenship of	5.202	2.961
citizens who join terrorist organisations.	(0.362)	(0.111)
This House regrets the EU's introduction of freedom of movement	4.260	2.798
	(0.361)	(0.101)
This House would suspend trade union powers and significantly relax labour protection	4.260	2.721
laws in times of economic crisis.	(0.360)	(0.104)
This House believes that causing deliberate harms to enemy civilians, by the	4.337	2.817
weaker side, is a justified tactic in asymmetrical warfare.	(0.346)	(0.112)
Observations	104	104
During periods of national housing shortages, this House would forcibly take	4.054	3.033
ownership of privately owned homes which are not lived in by their owners).	(0.358)	(0.113)
This House believes that states should aggressively fund geoengineering projects	4.152	3.352
instead of attempting to mitigate the effect of climate change.	(0.305)	(0.126)
This House regrets the decision to let the FARC (i.e. The Revolutionary	4.272	3.033
Armed Forces of Colombia -People's Army) run as a political party.	(0.442)	(0.103)
When tech companies own platform utilities and platform products,	3.739	2.835
this House would break them up.	(0.361)	(0.123)
Observations	92	92

Note: Column (1) reports the number of Points of Information, the event of a non-speaking debater standing up to challenge the speaker, received by each speaker. Column (2) reports the score, on a scale from 1 "Not heated at all" to 5 "Very heated" that the enumerator assigns to each speaker for her performance.

Table A.13: Pair-wise Correlation Between Measures of Debate Heat and Baseline Alignment

	POIs above median	Subjective heat scores above median	Baseline belief alignment
	(1)	(2)	(3)
POIs above median	1.000***		
	(0.000)		
Subjective heat scores above median	0.281***	1.000***	
	(0.002)	(0.000)	
Baseline belief alignment	0.184*	0.036	1.000***
	(0.051)	(0.702)	(0.000)
Observations	114	114	114

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Note: P-value for Statistical Significance in Parentheses The unit of observation for this analysis is a debate. The number of Points of Information and the subjective heat scores are aggregated at the debate room level, and for each of these aggregate measures we construct a binary indicator variable to denote, within each round, the debate rooms with aggregate score above median.

At the individual level, the first measure is a poor predictor of how heated the speaker is, because in fact the measure captures how heated the non-speaking debaters in the room are. Aggregating each of the two individual level measures at the debate room level allows us to obtain two outcomes that lend themselves to an interpretation in terms of heat. Table A.13 quantifies the correlation between the measures of heat of a debate: such correlation of 0.28 is substantial, but far from perfect. To complement the set of result on the correlation between alignment and persuasiveness, we show some evidence that the more debaters' beliefs turn out to be aligned with their persuasion goals, the more heated the debate turns out to be. This is interesting, because it suggests that debaters who truly believe in their position act more forcefully during the debate. Though, as shown in section 1.3, such additional energy does not translate into significantly better persuasion outcomes.

A.7 Robustness to Experimenter Demand Effects

When subjects of experimental work are able to infer the research hypotheses under investigation, we often worry that they may distort their reports to help the researchers prove their hypotheses. To reduce such concerns, one can raise the costs for subjects to distort their reports to conform to the researchers' hypotheses. This is what we achieve in our experiment by eliciting incentivized beliefs, and by asking subject to distribute monetary endowments between causes that generate real social returns.

By definition, for experimenter demand effects to potentially drive the results, it is necessary that subjects are able to infer the research hypotheses under investigation. To establish the extent to which they can, at the end of our study, we ask subjects of our experiment to write down in an open field text box what they thought the research was trying to demonstrate.

Table A.14: Categorization of Debaters' Response

(a) Having to argue for a given position alters the perception of empirical facts	0.227
	(0.032)
(b) Having to argue for a given position alters the perception of values	0.125
	(0.025)
(c) Having to argue for a given position makes individuals relatively more confident about the merit of their position	0.091
	(0.022)
(d) Positive correlation between private beliefs aligned with the persuasion goal and persuasiveness	0.142
	(0.026)
(e) Convergence of opinions through the debate	0.131
	(0.025)
(f) Other research questions	0.284
	(0.034)
(g) Overly generic answer	0.301
	(0.035)
Answered question	176
Left field blank	20
Observations	196

Notes: Open-field answers are categorized by a research assistant to be either an overly generic answer, or to reflect at least one of the research hypotheses (a) to (e) and possibly other potential research hypotheses. We report shares of respondents (and standard errors) in each category among the 90 percent of respondents who did not leave the open-field question unanswered.

The majority of subjects reported fairly sophisticated guesses.[4] In Table A.14 we report the result of our manual categorization of non-blank responses (90 percent of the sample). Among these, only 30 percent give an overly generic answer, while the rest seem to have in mind some concrete research hypotheses. The most frequent category is our residual category "Other research questions", that includes questions that were not part of our pre-registered hypotheses. Relatively frequently, subjects also seem to appreciate some reasonably close version of our primary research hypothesis of self-persuasion on facts.

Studies that try to bound the extent to which experimenter demand effects can explain experimental results, assess how sensitive results are to increasing awareness among subjects of the experimenters' research hypotheses De Quidt, Haushofer, and Roth, 2018. In the absence of such exogenous variation of awareness of research hypotheses, an imperfect but informative exercise that we can conduct is to provide evidence of how results change when we exclude from the test of a specific hypothesis the responses of subjects who were able to figure out that hypothesis. In Table A.15 we do exactly that to consolidate our self-persuasion results obtained by comparing belied, attitude, and confidence alignment with the persuasion goal. Reassuringly, we find that the magnitudes of the differences in all three outcomes between proposition and Opposition speakers, estimated for the subset of "unaware subjects", are very similar to the ones estimated in the full sample.

[4]Some responses were fairly accurate in capturing many of the research hypotheses (e.g. *"1. See how engaging with motion from a certain assigned point of view influences perception of facts in accordance to position in debate 2. how belief/being convinced of position in debate affects debaters persuasiveness (that's why you gave us scores on persuasion and rhetoric as well) –¿ How debating from assigned point of view affects opinion and how that affects performance in debate"*, some others completely miss the main hypotheses (e.g. *"Connection between knowledge and persuasivness? - Not sure, would love to find out!"*), and some others are overly generic (e.g. *"Game-theory"*).

Table A.15: Replication of Main Results Excluding Subjects Who Could Guess The Research Hypothesis at the End of the Tournament

	Beliefs aligned with proposition	Attitudes aligned with proposition	Confidence in proposition
	(1)	(2)	(3)
Speaker in proposition	0.235***	0.243*	4.325***
	(0.065)	(0.127)	(1.581)
Debater fixed effects	✓		✓
Round fixed effects	✓	✓	✓
Observations	698	779	813

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Note: Column (1) replicates analysis in column (1) of Table 1.3 excluding subjects who guessed the research hypothesis of self-persuasion on facts. Column (2) replicates analysis in column (1) of Table 1.4 excluding subjects who guessed the research hypothesis of self-persuasion on the values of social causes. Column (3) replicates analysis in column (1) of Table 1.5 excluding subjects who guessed the research hypothesis that debaters who be relatively more confident of the merits of their own position.

A.8 Mechanisms

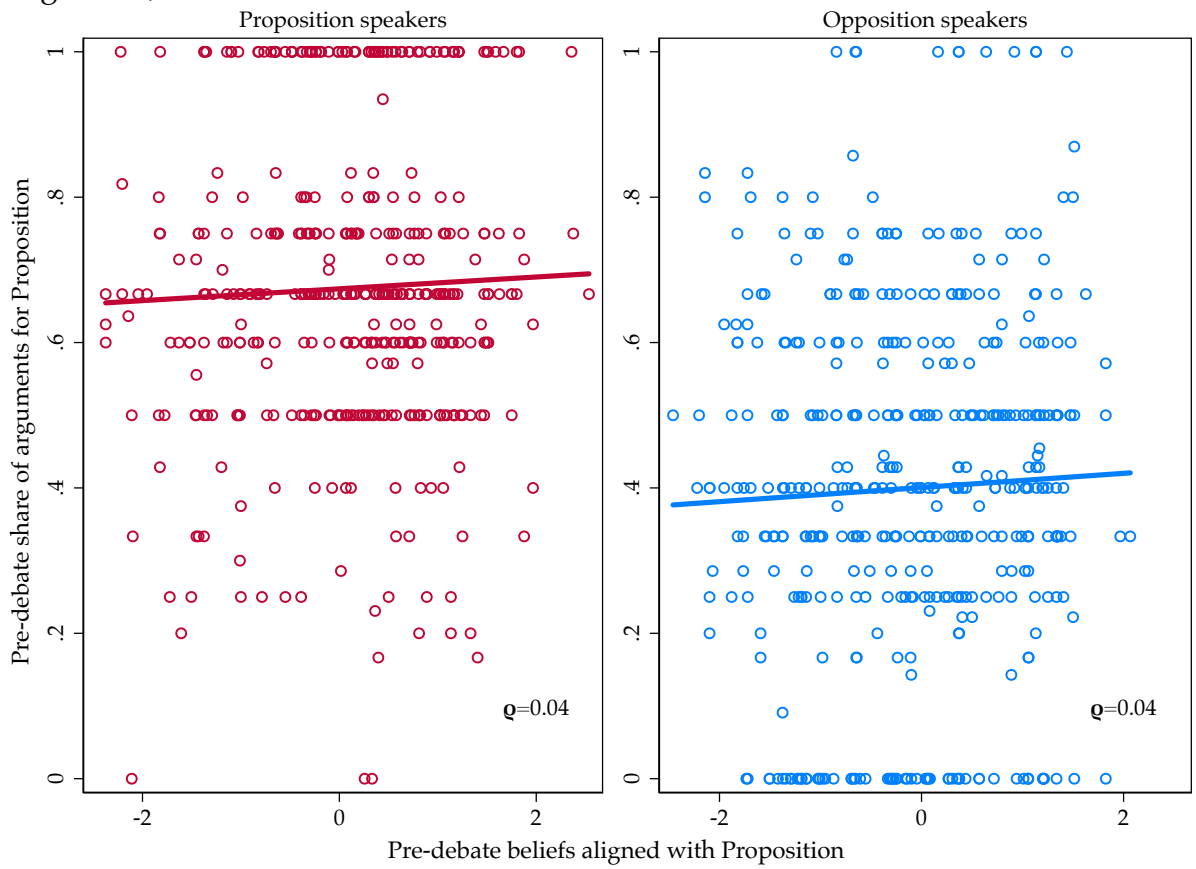
Our discussion proposes that persuasion goals can have both a direct effect on belief alignment due to strategic choice of beliefs and an indirect effect due to the cognitive constraints that generate bias when debaters sample an unbalanced set of arguments to prepare their speech. In a linear framework, such direct and indirect effects can be assessed through the following system of structural equations

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 T_i + \phi_1 X_i + \epsilon_{i1} \\ M_i &= \alpha_2 + \beta_2 T_i + \phi_2 X_i + \epsilon_{i2} \\ Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \phi_3 X_i + \epsilon_{i3} \end{aligned} \tag{A.1}$$

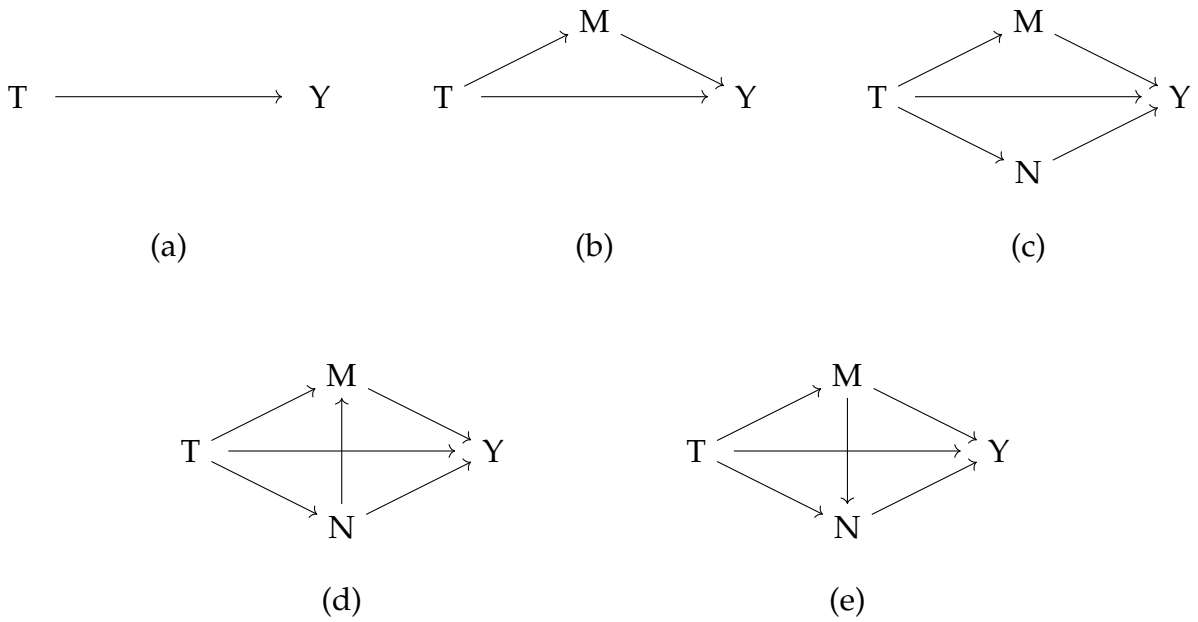
where standard notation is used for expositional purposes: Y_i is the outcome of interest, T_i is the treatment variable, M_i is the intermediate outcome measure after treatment that mediates the treatment effect, and X_i is a vector of controls. β_1 represents the average treatment effect (ATE), which includes both direct and indirect effects of the treatment on the main outcome of interest. If the structural equations are correctly specified, a *sequential ignorability* assumption allows to interpret $\gamma\beta_2$ as the *causal* indirect effect of T_i, mediated through M_i, on Y_i Imai, Keele, and Yamamoto, 2010.

Sequential ignorability requires that (i) conditional on X_i, the outcome and the mediator are distributed independently of the treatment, and (ii) conditional on T_i and X_i, the outcome is distributed independently of the mediator. Both conditions are fairly strong. Because our treatment assignment is randomized, the first condition is met by design. However, the second condition does not directly follow from random assignment, and is hard to test. If the second condition is met, we would expect that the outcome and the mediator are uncorrelated within treatment. Figure A.6 provides supporting evidence of the lack of such correlation.

Figure A.6: Correlation Between Share of proposition Arguments and Predebate Belief Alignment, Within Each Side of the Debate



In Figure A.7 we include diagrams that illustrate potential causal links between the treatment, mediating factors, and the outcome. Assuming sequential ignorability rules out causal links between mediators (sub-figures (d) and (e)), but allows for multiple downstream causal relationships from treatment, through mediators, to the outcome of interest (sub-figures (a) to (c)), so that by estimating $\gamma\beta_2$ from A.1 we could directly obtain a valid estimate of the causal effect of the treatment mediated through M_i.



Note: In (a), the outcome can only be affected directly by the treatment variable. In (b), the treatment affects both the outcome directly and an intermediate mediator; the mediator in turn affects the outcome. In (c), the treatment affects both the outcome directly and two intermediate mediators; both mediators in turn affect the outcome. In (d) and (e), the treatment affects both the outcome directly and two intermediate mediators; both mediators in turn affect the outcome, and mediators also affect one another.

Figure A.7: Diagrams Representing Possible Causal Mechanisms Between Treatment, Mediating Outcomes, and Main Outcome

In the potential outcome framework with binary treatment $t \in \{0, 1\}$ and one mediator it is straightforward to derive the causal mediated effect directly as a component of the average treatment effect $\tau_i = Y_i(1) - Y_i(0)$, which can be equivalently written as $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$. With some algebra, it is simple to obtain that

$$2[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] = \overbrace{Y_i(1, M_i(1)) - Y_i(1, M_i(0))}^{\delta_i(1)} + \overbrace{Y_i(0, M_i(1)) - Y_i(0, M_i(0))}^{\delta_i(0)} +$$
$$+ \overbrace{Y_i(1, M_i(1)) - Y_i(0, M_i(1))}^{\zeta_i(1)} + \overbrace{Y_i(1, M_i(0)) - Y_i(0, M_i(0))}^{\zeta_i(0)}$$

where $\delta(t)$ defines the indirect effect of the treatment in treatment t, and $\zeta_i(t)$ defines the direct effect of the treatment holding constant the level of the mediator at the treatment t level. When $\delta_i(t) = \delta_i$ and $\zeta_i(t) = \zeta_i$ for any t, there is no interaction between treatment and mediator, and the ATE can simply be expressed as $\tau_i = \delta_i + \zeta_i$, yielding a simple decomposition of the ATE in average causal mediated effect (ACME) and average direct effect (ADE).

To identify the ACME of persuasion goals on belief alignment with proposition b_i through the share of proposition arguments considered during preparation period s_i, we estimate the following random effects models with standard errors clustered at the team level

$$\textit{Model 1:} \qquad b_{i,m} = \alpha_1 + \beta_1 proposition_{i,m} + \phi_1 X_i + \epsilon_{i1,m}$$

$$\textit{Model 2:} \qquad s_{i,m} = \alpha_2 + \beta_2 proposition_{i,m} + \phi_2 X_i + \epsilon_{i2,m}$$

$$\textit{Model 3:} \qquad b_{i,m} = \alpha_3 + \beta_3 proposition_{i,m} + \gamma s_{i,m} + \phi_3 X_i + \epsilon_{i3,m}$$

and use sampling distributions of the parameter estimates from *model 1* to simulate potential outcomes $b_{i,m}(proposition_{i,m} = 1)$ and $b_{i,m}(proposition_{i,m} = 0)$, from *model 2* to simulate potential outcomes $s_{i,m}(proposition_{i,m} = 1)$ and $s_{i,m}(proposition_{i,m} = 0)$, and from *model 3* to simulate potential outcomes $b_{i,m}(1, s_{i,m}(1))$, $b_{i,m}(0, s_{i,m}(1))$, $b_{i,m}(1, s_{i,m}(0))$, and $b_{i,m}(0, s_{i,m}(0))$. Table 1.6 in the main text reports the results from this exercise.

A.9 Surveys

A.9.1 General instructions

A two-page general instructions document includes relevant information for answering the surveys throughout the tournament. In particular this explains how belief elicitations are incentivized using the Quadratic Scoring Rule for binarized outcomes (Harrison, Martínez-Correa, and Swarthout, 2014), how charitable allocations are paid out, and general payment procedures. All subjects are given 10 minutes to carefully read these general instructions right before the baseline survey begins. To make sure that procedures are adequately understood, if subjects miss their opportunity to read the general instructions we exclude them from the study.[5] The original content of these instructions is provided below.

. . . . . .

General Instructions

Please read the following instructions carefully and keep them in mind, as they contain information that is relevant for the surveys we will ask you to complete during the next two days. We kindly ask you to use the time allocated to each survey to focus exclusively on answering the questions in front of you; throughout these times no information regarding the debates will be provided. Please answer each question carefully, don't use your phone and don't interact with others. Our instructions are never deceptive. All of your answers are treated confidentially and used for research purposes only.

Assessing factual statements

Spread across the various surveys, there are 34 questions that are marked by an "$", for which you can earn money. After you completed the last survey, we will pay you

[5] They are allowed to answer the surveys, but their data is discarded.

based on one randomly selected answer. While you will get paid for only one of your answers, every question might be the one that counts.

Questions marked by an "$" ask you to state the likelihood (in percent) that a given statement is true. Most such statements are designed to assess your factual knowledge. There will be no trick questions. Moreover, all sources we refer to actually exist and are of high quality, but the actual fact may be either true or not true. As an example, consider the following statement.

According to Eurostat, more than 30 percent of live births in Germany in 2016 were outside of marriage.

This statement is true if Eurostat indeed reported this finding. It is false if Eurostat reported a different finding. You will be asked to provide your belief as to how likely you think it is that this statement is true. If this answer is selected for payment, you will earn either 30 euros or nothing. The procedure that determines how likely it is that you win the 30 euros assures that the closer you are to the correct answer (either 0 or 100 percent), the higher is your probability of winning the money.

Moreover, the procedure assures that you maximize your chance of winning money by stating your true belief (between 0 and 100 percent). So if you are almost certain that a given statement is true, then you should state a belief that is very high. If you are almost certain that a given statement is false, then you should state a belief that is very low. If you are completely uncertain, you maximize your chance of winning by stating a belief that is close to 50 percent.

The Procedure Box below provides more comprehensive information about the exact payment mechanism. But note that it is not important that you understand the procedure in detail. What matters is that you know that you maximize your probability of winning when you report your true belief - if you under- or overstate your belief, you will reduce your chance of winning the 30 euros.

Donating to Charities

For some questions in the survey, you will be able to allocate monetary endowments

between different charities. This is money that we make available from our budget for you to allocate, according to your preferences, to charities that have different missions. One of the allocations you make will be selected at random and we will transfer the money to the relevant charities. While we will implement only one of your allocations, every allocation might be the one that counts.

The surveys will also feature further questions that allow you to earn more money for yourself. The instructions for these questions are simple and will be provided above the relevant question.

Procedure Box

How a given answer maps into your chance of winning 30 euros is based on a formula. This formula is designed to make sure that you maximize your chance of winning if you report your true belief that a given statement is true.

Suppose that the correct answer is given by R, which is equal to 1 if the statement is true and 0 if the statement is false. The variable r is your report–the likelihood that you attribute to the statement being true (from 0 to 100 percent). The winning probability for the prize is then given by:

$$\text{winning probability} = 100 - 100 \times (R - r/100)^2$$

Example: Suppose again that you are tasked with assessing the following statement: *According to Eurostat, more than 30 percent of live births in Germany in 2016 were outside of marriage*. And suppose that your belief that the statement is true is 63 percent. The following table shows your winning probability based on the formula. The columns represent a number of hypothetical answers you may give. As you can see, you maximize your chance of winning by reporting your true belief.

	Report 1	Report 2	**Report 3**	Report 4
Hypothetical report	22	35	**63**	89
Expected winning probability if **your belief** that the statement is true is **63%**	59.9%	68.9%	**76.7%**	69.9%

Payment

On Sunday, we will pay out your earnings in cash. To determine your earnings for the assessment of factual statements, we first randomly draw the question that is relevant for your payment. We then determine your winning probability based on the true answer and your reported answer. Finally, a computer program constructs a virtual urn with only white and black balls, where the share of white balls equals your winning probability. If the computer then draws a white ball from the urn, then you will win the 30-euro prize. This is a fair and transparent procedure to pay you the prize with

the winning probability you have earned based on the quality of your answers.

If the question that is drawn for payment is from a round that you missed, then there will be no new draw and you will not earn any money for this type of question. If you would like us to send you receipts of the charity donation based on your choice, then please leave us your email address when you collect your payment.

A.9.2 General remarks

We take several steps to collect high quality data in a confidential manner.

First, all surveys that debaters fill out begin with a cover page containing brief instructions to (i) inform subjects how much time they have to complete the survey, and (ii) remind subjects of the procedure to collect incentive compatible beliefs. The cover page does not contain any question, and enumerators are instructed to not turn the cover page after surveys are filled out and read the answers provided by debaters.

Second, each survey is linked to the individual who filled it through a personal identifier. Debaters are assigned S### IDs, Judges J## IDs, and Enumerators E## IDs. These IDs allow data to be collected and payments to be carried out confidentially. We ask debaters to enter their S IDs on the cover page of each of their surveys.

Every study participant (debaters, judges, and enumerators) wears a name tag that includes their ID. Before collecting the survey, enumerators double-check that the S ID entered by each debater on the cover page of their survey matches the one on the name tag.

A.9.3 Baseline survey

A 25-minute baseline survey includes the following items:

- Age (open field, suggested to provide a numeric answer).
- Gender (open field).
- Nationality (open field).

- Political ideology scale: "*In politics people sometimes talk of "left" and "right". Where would you place yourself on this scale, where 0 means the left and 10 means the right?*" (check box).

- Years actively debating on a regular basis. Options: "Less than a year", "1 to 2 years", "3 to 4 years", "At least 5 years". (check box)

- Times debater got to semifinals in Open/IV tournaments (open field).

- "*What do you think makes a good debater*". Options: "Choosing arguments strategically", "Confidence in own position", "Debating experience", "Factual knowledge", "Eloquence" (ranking).

- Incentivized belief elicitation on fifteen factual statements: for each such statement subjects state how likely it is that the fact is true (open field, suggested to provide a numeric answer from 0 to 100).

- "*Did you take part as a speaker at the Munich Research Open 2019?*". Options: "Yes", "No" (check box).[6]

A key component of this survey was to gather beliefs at baseline regarding the motions that subjects were going to debate. At the same time, we had to be careful in not revealing, through our questions, the motion of the debates – which are meant to be secret. To obfuscate the relation of these belief elicitations and the motions we elicit beliefs over whether 15 factual statements are true: 5 such statements relate to the in-round motions, 7 are decoy questions, and 3 are control questions.[7] For each team of debaters, control questions are drawn from a pool of 6 questions, and the questions that were not selected for the baseline survey are then included in the endline survey. Comparing responses to the control questions at baseline and endline by different debaters helps uncover to what extent debaters discuss the contents of the surveys among themselves.

[6]Only in Rotterdam.

[7]In Rotterdam, 4 statements relate to the in-round motions, and 8 are decoy questions.

Decoy questions are designed to look like they could relate to plausible motions for debate. Control questions are facts that not necessarily relate to typical debate topics.

For each motion, we devise multiple factual statements that we phrase as binary states to capture alignment of beliefs with the persuasion goal. Any given question may not have a tight enough link to the motion in debaters' minds or give rise to a high degree of certainty in debaters' beliefs and may therefore be ill-suited to pick up a treatment effect. To diversify this risk, we come up with 4 questions (A, B, C, D) for each motion and administer them as illustrated in the table below: at baseline, debaters are asked either about fact A or B; predebate, debaters are asked either about fact D or C; postdebate debaters are asked either about fact B and C or A and D.

This approach also ensures that (i) no debater is asked the same question twice, and (ii) we protect the baseline and predebate belief elicitations from any potential information spillovers.

Timing:	Beginning of Day 1	Day 1 or Day 2	
	Baseline	Predebate	Postdebate
Subgroup 1	A	D	B, C
Subgroup 2	B	C	A, D

A.9.4 Predebate survey

This 5 minute survey is handed out before each debate begins and after the preparation time. It includes:

- Incentivized belief elicitation on two factual statements: for each such statement subjects state how likely it is that the fact is true (open field, suggested to provide a numeric answer from 0 to 100).
- Choose one of 9 monetary allocations, along a concave budget, between a baseline charity (either Oxfam or Opportunity International) and a charity aligned with one of the sides represented in the debate. For an illustration see Figure A.8.

- Questions on the number of arguments considered during preparation time in favor of the proposition:

  i *How many good arguments did you come up with during the preparation time in favor of the proposition?* (open field, suggested to provide a numeric answer)

  ii *How many of these arguments would you consider to be very strong?* (open field, suggested to provide a numeric answer between zero and the answer to the previous question)

- Questions on the number of arguments considered during preparation time against the proposition:

  i *How many good arguments did you come up with during the preparation time against the proposition?* (open field, suggested to provide a numeric answer).

  ii *How many of these arguments would you consider to be very strong?* (open field, suggested to provide a numeric answer between zero and the answer to the previous question).

Figure A.8: Illustration of charitable donations allocation question

Below you see nine potential ways in which you could allocate charitable donations—that are paid by us on your behalf—between two charitable organizations: Oxfam and The Planetary Society .

Oxfam is a major nonprofit group with an extensive collection of operations. Oxfam's programs address the structural causes of poverty and related injustice and work primarily through local accountable organizations, seeking to enhance their effectiveness

The Planetary Society is the world's largest and most influential non-profit space organization. The society advocates for space and planetary science funding in government, invests in inspiring educational programs, and funds groundbreaking space science and technology

How would you like to allocate these donations? (check only one box)

Choose one option	☐	☐	☐	☐	☐	☐	☐	☐	☐	
You want to give	0.0	1.3	2.5	3.7	5.0	5.9	6.4	6.7	7.0	euro to Oxfam
and	7.0	6.7	6.4	5.9	5.0	3.7	2.5	1.3	0.0	euro to the Planetary Society
A total of	7.0	8.0	8.9	9.6	10.0	9.6	8.9	8.0	7.0	*euro goes to charity*

Both factual statements are meant to capture whether beliefs are aligned with the motion after the debate. The first statement features a real-world fact. The second

statement elicits confidence in the arguments of the proposition side of the debate by asking:

Statement: Excluding the debate happening in this room, in at least half of the parallel debates of this round, one of the two teams on the Government side of this motion will rank 1st.

Q2$: How likely do you think it is that the above statement is true? ___% (write a number from 0 to 100)

For each motion, we select two charities that we expect to be either positively or negatively aligned. We randomly determine which of these two charities features in the predebate survey. The other charity features in the postdebate survey. In Rotterdam, the baseline charity is always Opportunity International, whereas in Munich we also randomize between Oxfam and Opportunity International to be the baseline charity.

A.9.5 Postdebate survey

This 5 minute survey is handed out right after each debate. It includes:

- Incentivized belief elicitation on two factual statements: for each such statement subjects state how likely it is that the fact is true (open field, suggested to provide a numeric answer from 0 to 100).
- Subjective ranking of team performance in the debate.
- Choose one of 9 monetary allocations, along a concave budget, between a baseline charity (either Oxfam or Opportunity International) and a charity aligned with one of the sides represented in the debate. For an illustration see Figure A.8.

Both factual statements are meant to capture whether beliefs about real-world facts are aligned with the motion after the debate.

For each motion, we select two charities that we expect to be either positively or negatively aligned. We randomly determine which of these two charities features in the predebate survey. The other charity features in the postdebate survey. In Rotterdam, the baseline charity is always Opportunity International, whereas in Munich

we also randomize between Oxfam and Opportunity International to be the baseline charity.

A.9.6 Endline survey

This 20-minute survey takes place right after the fifth (fourth in Rotterdam) round of debates.

It includes:

- A question that we use to assess how debaters think that beliefs about facts that we ask and charities they can donate to relate to alignment with the motions. An illustration of the precise wording of this question is provided in Figure A.9.
- Incentivized belief elicitation on four factual statements: for each such statement subjects state how likely it is that the fact is true (open field, suggested to provide a numeric answer from 0 to 100).
- Open text box in which subjects are asked to tell us what they think the research was about.[8]

Three of the four factual statements are control questions of the kind included in the baseline survey. One fact pertains the performance of two actual debaters in the Munich Research Open, and had a longer preamble than other belief elicitation questions:

The next question is about the performance of two actual debaters in a different tournament: the Munich Research Open that took place two weeks ago[9]*. We will call them debater A and debater B. Both debaters were representing the Government in the motion that "THBT governments should stop funding scientific programmes that have no immediate benefit for humankind (such as space travel and exploration, human cloning)", but they gave different responses to the factual question in the predebate survey:*

[8] We felt that the alignment question was revealing too much of what the study was about, so to get a better sense of whether subjects understood what hypotheses were being tested with the data collected in predebate and postdebate surveys, in Rotterdam, we decided to move this question to the last postdebate survey.

[9] In Rotterdam. In Munich, the orange text is replaced by "this tournament".

Debater A believed that the statement "More than 10 of the following 15 innovations are a consequence of inventions made in the pursuit of space travel: camera phones, scratch resistant lenses, electric light, CAT scans, LEDs, land mine removal, athletic shoes, penicillin, water purification systems, the internet, home insulations, wireless headsets, baby formula, portable computers" was true with 75% chance. Debater B believed that the same statement was true with 10% chance.

We asked judges to provide a broad measure of each debaters' persuasiveness. Now consider the following statement.

Statement: Debater A obtained a higher persuasiveness score than Debater B in the relevant debate.

Q6$: How likely do you think it is that the above statement is true? ___ % (write a number from 0 to 100)

Figure A.9: Example of Aligment Question in the Endline Survey

Q1: For this question, you can earn up to 5 euros. Consider the following motion, which was debated during the event:

"During periods of national housing shortages, this House would forcibly take ownership of privately owned homes which are not lived in by their owners".

Now consider someone who is strongly in favor of this motion, i.e. someone whose personal views are strongly aligned with the motion. For each statement in the table below, please indicate whether such a person, who is strongly aligned with the motion, is more likely to believe that the statement is true or more likely to believe that the statement is false. For each statement (each row of the table), give your answer by entering one (and only one) cross in the appropriate box.

We will randomly select one of these six statements and pay you based on your selection as follows: You will earn 5 euros for sure if your response is the same as the response that is selected most frequently by all other participants answering the same question.

	Someone aligned with the motion is…		
	… more likely to believe that this statement is **true**	… more likely to believe that this statement is **false**	.. equally likely to believe that this statement is **true or false**
Statement 1: According to the English Housing Survey, the number of second homes in the UK more than doubled between 1995 and 2013			
Statement 2: Under current UK regulation, squatters who live in and maintain unoccupied buildings enjoy protection under the law and can never be evicted without a court order			
Statement 3: According to an academic study published this year, over 5 percent of properties in England and Wales are low-use properties, defined as a property that is not registered as the primary residence of any individual			
Statement 4: According to research by the newspaper the Independent in 2018, more than one third of new-build luxury apartments and houses in Central London lies empty			
Statement 5: Action on Empty Homes* is an NGO supporting a cause that is especially important.			
Statement 6: The Land Is Ours** is an NGO supporting a cause that is especially important.			

*Action on Empty Homes is a UK NGO campaigning for more empty homes to be brought into use for people in housing need. It raises awareness of the waste of long-term empty homes and campaigns for changes to national policy to bring more homes into use..

**The Land Is Ours campaigns peacefully for access to the land, its resources, and the decision-making processes affecting them. Among other things, it advocates 'Use It Or Lose It' programme where empty buildings are forfeit or put on a tax escalator, where the owner can lose title after one year.

A.9.7 Judge survey

Judges are asked to independently provide individual scores of each debater's overall persuasiveness before filling out the shared score sheet with other judges.

Judges are asked to provide a broad persuasiveness score, on a scale from 1 to 10 where 1 is "Not at all persuasive" and 10 "Extremely persuasive". The original instructions given to judges on how to answer and interpret this question are provided below:

Without discussing with the other judges, please evaluate the persuasiveness of each debater. We consider a debater persuasive, if she would do well at convincing a general audience of her position. Therefore, please provide a broad measure of persuasiveness that captures the quality of arguments as well as speaking ability, body language and any other attribute that makes a speech persuasive to a general audience.

To ensure that the judges provided independent persuasiveness scores, we asked them to fill out these surveys during the debate. Judges on the panel painstakingly take notes of each speech and generally do not interact with each other during the debate. We collected the surveys before any deliberation of the panel took place.

A.9.8 Enumerator survey

A survey that the enumerator answers during the debate includes the following items:

- A count of the times not speaking debaters try to interrupt the speaker (through Points of Information).
- A subjective rating of how heated each debaters' argumentation is coming across (on a scale from 1 to 5).[10]
- For each of the four facts related to the motion over which we elicit debaters beliefs, and for both the motion related charities, note whether these were mentioned during the debate.

A.9.9 Ballot

The ballot is the official module that debating tournament have panels of judges fill out to evaluate a debate. This form includes:

- Name and position of each team in the debate
- Ranking of the four teams in the debate (from First to Fourth, with no possibility for ties)

[10]Enumerators were instructed to write down this score for each debater at the end of the speech. They could however revise this score for debaters that acted particularly heatedly during other debaters' speeches.

- Individual speaker scores (on a scale from 50 to 100)

After a debate is over, speakers leave the room to let judges on the panel privately discuss the performance of each debater. This discussion takes approximately 15 minutes during which the arguments presented by each debater are technically analyzed. A technical analysis is particularly relevant to the assignment of individual speaker scores, which are supposed to be assigned an a objective scale that applies to any British Parliamentary performance.[11] The ballot is filled out at the end of this discussion.

[11] An example of such scale can be found at https://debate.uvm.edu.

A.10 Motion Facts and Charities

Table A.16: Decoy and Control Belief Elicitations for Baseline Survey in Munich

Fact
Decoy questions
1. The US has more nuclear weapons than any other country.
2. A paper recently published in a leading economics journals finds that the decriminalization of prostitution in Rhode Island in 2003 caused reported rape offences to fall by over 20%.
3. A recent randomized controlled trial with almost 3000 social media users finds that individuals that are paid to stay off of Facebook for four weeks watch more TV and are less informed about current events.
4. As measured by the Eurobarometer survey, a majority of Europeans are not interested in receiving information about treatment conditions of farm animals.
5. According to a review published in a prominent public health journal in 2011, nutrition labels are a cost effective intervention to promote healthier diets.
6. A paper published in a leading economics journal in 2009 finds that violent crime increases on days with larger theater audiences for violent movies.
7. According to a 2019 review study in a prominent scientific journal, the well-being of teenagers has a stronger relation with having regular breakfast habits than with the use of digital technologies.
Control questions
1. The corporate income tax is higher in the US than in Finland.
2. In France, government spending was over half of GDP in 2017.
3. More than half of children in the United States were overweight or obese as of 2014 (BMI of 25 or greater).
4. Less than 30% of all Nobel prizes in Chemistry were awarded to U.S. citizens.
5. The PISA is a worldwide exam administered every three years that measures science, reading and math skills of 15-year-olds. In 2015, at least 4 Asian countries were in the top 10 in each category of the exam.
6. According to the UNESCO, the global literacy rate is under 90%.

Note: All decoy questions are included in the baseline survey. For each subject we randomize whether only the first three control question or the last three control questions are included in the baseline survey; the other three questions are included in the endline survey.

Table A.17: Alignment of facts with motions in Munich

	Fact	Alignment predicted by: Authors	Alignment predicted by: Debaters
	This House believes that governments should stop funding scientific programs that have no immediate benefit for humankind (such as space travel and exploration, human cloning)		
Motion 1	A. The European Space Agency's annual budget exceeds 4% of the EU budget	proposition	proposition (65%)
	B. More than six Europeans out of ten agree that space technologies have a role to play in avoiding threats like asteroids, comets, and space debris collisions	Opposition	Opposition (50%)
	C. More than 10 of the following 15 innovations are a consequence of inventions made in the pursuit of space travel: camera phones, scratch resistant lenses, electric light, CAT scans, LEDs, land mine removal, athletic shoes, penicillin, water purification systems, the internet, home insulation, wireless headsets, baby formula, portable computers	Opposition	Opposition (70%)
	D. A study in a leading economics journal in 1998 finds that the majority of R&D spending by the US government goes into wages for scientists, which in turn does little to increase the number of scientists	proposition	proposition (65%)
	This House believes that Western States should permanently revoke the citizenship of citizens who join terrorist organisations		
Motion 2	A. From 2015 to 2017 there were more than 50 separate Islamic terrorist attacks in the EU	proposition	proposition (90%)
	B. According to the UN's Basic Human Right's Reference Guide on the right to a fair trial and due process in the context of countering terrorism, the citizenship of people suspected or proven of having been part of terrorist organisations may never be revoked	Opposition	Opposition (60%)
	C. In Germany, the law governing citizenship already permits to strip those with dual citizenship of their German citizenship if they join a foreign army	proposition	proposition (70%)
	D. The two main perpetrators of the 2015 attacks on the offices of the satirical newspaper Charlie Hebdo had fought with IS in Syria prior to the attack	proposition	proposition (85%)
	This House regrets the EU's introduction of freedom of movement		
Motion 3	A. In a much-cited academic article from 2012, researchers from University College London found that immigration increased wages in the UK, both at the bottom and at the top of the income distribution	Opposition	Opposition (65%)
	B. More than 35% of UK citizens interviewed for the Eurobarometer in 2018 think that the Schengen Area has more disadvantages than advantages for the UK	proposition	proposition (80%)
	C. According to a 2018 paper by researchers from the University of Munich, emigration within Europe positively contributes to innovation in source countries, i.e. countries people emigrate from	Opposition	Opposition (70%)
	D. Less than half of Europeans agree that integration of immigrants has been a success in their local area, city or country	proposition	proposition (65%)
	This House would suspend trade union powers and significantly relax labour protection laws in times of economic crisis		
Motion 4	A. A 2015 survey by the Employment Policies Institute shows that a majority of economists thinks that a USD 15 minimum wage will reduce the number of jobs	proposition	proposition (73%)
	B. According to a study of 21 Eastern European economies published in 2017, members of labor unions are less likely to lose their job during an economic crisis	Opposition	Opposition (46%)
	C. The fraction of the U.S. population that approves of labor unions dropped by more than one third from the mid-50s to 2009, according to the Gallup poll	proposition	proposition (86%)
	D. In a 2005 study of OECD countries, economists from Harvard University and the University of Bonn find that greater labor market flexibility (e.g. due to weaker trade unions) is associated with greater employment	proposition	proposition (59%)
	This House believes that causing deliberate harms to enemy civilians, by the weaker side, is a justified tactic in asymmetrical warfare		
Motion 5	A. Research on the psychological effect of violence against Israeli civilians shows that such violence caused a hardening of attitudes, stronger opposition to political reconciliation with perpetrators, and an increase in support for counter-terrorist measures	proposition	proposition (45%)
	B. Nelson Mandela, who was awarded the Nobel Peace Prize for facilitating South Africa's peaceful transition into democracy, was also the co-founder of the violent paramilitary wing of the African National Congress and was classified as a terrorist by the US until 2008	proposition	proposition (73%)
	C. Willingness to resort to violence was an integral part of the Birmingham campaign, which is widely credited with bringing about civil rights and desegregation in the United States of the 1960s	proposition	proposition (73%)
	D. Studies in political science consistently find that that rebel groups that use indiscriminate violence against civilians are more likely to achieve political goals	proposition	proposition (77%)

Table A.18: Alignment of charitable causes with motions in Munich

	Charitable cause	Alignment predicted by: Authors	Alignment predicted by: Debaters
	This House believes that governments should stop funding scientific programs that have no immediate benefit for humankind (such as space travel and exploration, human cloning)		
Motion 1	**The International Space University** develops the future leaders of the world space community. It encourages the innovative development of space for peaceful purposes: to improve life on Earth and advance humanity into space	Opposition	Opposition (80%)
Motion 1	**The Planetary Society** is the world's largest and most influential non-profit space organization. The society advocates for space and planetary science funding in government, invests in inspiring educational programs, and funds groundbreaking space science and technology	Opposition	Opposition (65%)
	This House believes that Western States should permanently revoke the citizenship of citizens who join terrorist organisations		
Motion 2	**The Active Change Foundation** is based in the UK and provides a holistic approach to neutralising extremism and violence on both an individual and community level. Its chief executive is an outspoken critic of those actors within the UK that favor stripping individuals of their citizenship for being involved with terrorist organisations	Opposition	Opposition (65%)
Motion 2	Human Rights Watch defends the rights of people worldwide. It scrupulously investigates abuses, exposes the facts widely, and pressures those with power to respect rights and secure justice. It has been a vocal defender of the right to citizenship for all people	Opposition	No relation (50%)
	This House regrets the EU's introduction of freedom of movement		
Motion 3	**The European Movement UK** is a grass-roots, independent, pro-European organisation. One of its main goals is to safeguard the freedom of movement made possible by membership of the EU, both for UK citizens who want to travel and work abroad and for citizens of other EU countries who want to come to the UK to work and to live	Opposition	Opposition (85%)
Motion 3	**ACT4FreeMovement** stands for Advocacy, Complaints, Trainings for Freedom of Movement. The organization campaigns for freedom of movement with EU citizens. The goal is to increase the capacity of EU citizens to effectively secure access to and knowledge of their rights, as well as build public awareness and political support for mobile citizen rights	Opposition	Opposition (75%)
	This House would suspend trade union powers and significantly relax labour protection laws in times of economic crisis		
Motion 4	**The European Trade Union Confederation** speaks with a single voice on behalf of European workers to have a stronger say in EU decision-making. It aims to ensure that the EU is not just an economic union but also a Social Europe, where improving the well-being of workers and their families is an equally important priority	Opposition	Opposition (64%)
Motion 4	**The Living Wage Foundation** is a campaigning organization in the United Kingdom, which aims to persuade employers to pay a Living Wage, an independently calculated and recommended minimum wage to cover workers' basic needs	Opposition	No relation (50%)
	This House believes that causing deliberate harms to enemy civilians, by the weaker side, is a justified tactic in asymmetrical warfare		
Motion 5	**The Israel Trauma Center** for Victims of Terror and War is an apolitical organization providing multidisciplinary treatment and support to direct and indirect victims of trauma due to terror and war in Israel	proposition	Opposition (50%)
Motion 5	**Muslim Aid** is an Islamic Charity, which has been actively working in Gaza since 2006. It helps vulnerable people to obtain essentials like food and medical supplies, which are scarce as importing and exporting has been made difficult	proposition	No relation (45%)

Table A.19: Decoy and Control Belief Elicitations for Baseline Survey in Rotterdam

Fact
Decoy questions
1. In 2016, from an estimated pre-war population of 22 million the UN estimates that more than 10 million people have been displaced internally as well as abroad.
2. A paper recently published in a leading economics journals finds that withdrawing legal access to cannabis improves academic performance of foreign university students affected by the policy in the Netherlands.
3. A recent The Lancet article finds that from the 15.6 million abortions that took place in India in 2015 over 10 percent were carried out outside of health facilities using unsafe methods.
4. A paper published in a leading economic journal estimates that juvenile incarceration in the US increases incarceration rates of individuals when they become adults.
5. A large representative survey published in a leading economic journal this year finds that over 30% of Americans would support a policy that allows recipients of kidney transplants to compensate living donors 100'000 USD in cash.
6. In the United States, more than half of all guns are sold without background checks.
7. A paper published in a leading economics journal in 2009 finds that violent crime increases on days with larger theater audiences for violent movies.
8. According to a 2019 review study in a prominent scientific journal, the well-being of teenagers has a stronger relation with having regular breakfast habits than with the use of digital technologies.
Control questions
1. Americans drink more alcohol per person than Europeans.
2. More than 30% of Europeans are smokers.
3. The PISA is a worldwide exam administered every three years that measures science, reading and math skills of 15-year-olds. In 2015, at least 4 Asian countries were in the top 10 in each category of the exam.
4. According to the 2015 Eurobarometer, more than 50% of Europeans feel that diversity is sufficiently reflected in the media in terms of religion or beliefs.
5. According to the 2015 Eurobarometer, more than 90% of Europeans say that they would feel comfortable with having a woman in the highest elected position in their country.
6. According to the UNESCO, the global literacy rate is under 90%.

Note: All decoy questions are included in the baseline survey. We included in the survey one more decoy question than we had in Munich to balance for the one fewer motion question (the experiment in Rotterdam covers only four rounds of debate). For each subject we randomize whether only the first three control question or the last three control questions are included in the baseline survey; the other three questions are included in the endline survey.

Table A.20: Alignment of facts with motions in Rotterdam

	Fact	Alignment predicted by: Authors	Alignment predicted by: Debaters
	During periods of national housing shortages, this House would forcibly take ownership of privately owned homes which are not lived in by their owners)		
Motion 1	A. According to the English Housing Survey, the number of second homes in the UK more than doubled between 1995 and 2013	proposition	proposition (72%)
Motion 1	B. Under current UK regulation, squatters who live in and maintain unoccupied buildings enjoy protection under the law and can never be evicted without a court order	Opposition	Opposition (57%)
Motion 1	C. According to an academic study published this year, over 5 percent of properties in England and Wales are low-use properties, defined as a property that is not registered as the primary residence of any individual	proposition	proposition (74%)
Motion 1	D. According to research by the newspaper the Independent in 2018, more than one third of new-build luxury apartments and houses in Central London lies empty	proposition	proposition (74%)
	This House believes that states should aggressively fund geoengineering projects instead of attempting to mitigate the effect of climate change		
Motion 2	A. Germany's experience with renewable energy promotion (i.e. its Renewable Energy Sources Act (EEG)) is often used as a model to be replicated elsewhere. Instead, a widely cited scientific study from 2010 argues that the German government's support of renewables has resulted in massive expenditures (annual feed-in tariffs of over 7 billion euros) that show little long-term promise for stimulating the economy, protecting the environment, or increasing energy security	proposition	proposition (74%)
Motion 2	B. According to recent data from the Climate Action Tracker, more than one third of the surveyed countries are well on track to meet the CO2 emission targets they imposed on themselves under the Paris agreement	Opposition	Opposition (50%)
Motion 2	C. Even the US, which has not supported recent global efforts to fight climate change by means of reducing CO2 emissions, has been enthusiastic in its support for geoengineering projects, as evidenced by its support for the U.N. resolution on geoengineering	proposition	proposition (70%)
Motion 2	D. A 2018 study by two prominent economists from MIT argues that increased investments in geoengineering may also increase efforts to improve clean energy technologies	proposition	proposition (78%)
	This House regrets the decision to let the FARC (i.e. The Revolutionary Armed Forces of Colombia - People's Army) run as a political party.		
Motion 3	A. Shortly after the 2016 peace deal with FARC, Colombia has been experiencing a resurgence of violence. The number of homicides is up by more than 7% in 2018 compared to the previous year	proposition	proposition (92%)
Motion 3	B. In 2016, the Nobel peace prize was jointly awarded to Colombian president Santos and the leader of FARC, Rodrigo Londoño, for their "resolute efforts to bring the country's more than 50-year-long civil war to an end"	Opposition	Opposition (60%)
Motion 3	C. In March 2017, the Colombian government reported that more than 25% of the estimated 6'900 FARC fighters refused to disarm	proposition	proposition (77%)
Motion 3	D. Towards the end of the peace deal negotiations between the Colombian government and FARC, NGOs like Amnesty International and Human Rights Watch as well as the Colombian Conservative party criticized the peace deal for being too lenient on perpetrators of human rights violations	proposition	proposition (90%)
	When tech companies own platform utilities and platform products, this House would break them up.		
Motion 4	A. According to a 2018 survey from the Pew Research Center, over 50% of Americans believe that major tech companies have too much power and influence in today's economy	proposition	proposition (88%)
Motion 4	B. The UK government's digital competition expert panel, chaired by Professor Furman who was chief economic advisor in Obama's presidency, issued a report just two weeks ago rejecting the widely held view that "digital platforms are natural monopolies where only a small number of firms can succeed"	Opposition	Opposition (77%)
Motion 4	C. According to a 2018 survey from the Pew Research Center, over 60% of Americans believe that major tech companies should be more regulated than they currently are	proposition	proposition (54%)
Motion 4	D. A 2018 survey of 1200 sellers on the Amazon platform, conducted by the independent market research firm Feedvisor, finds that over 40% of private sellers on Amazon fear that the company will take away their seller privileges and over 60% of them fear Amazon competing directly with them	proposition	proposition (92%)

Table A.21: Alignment of charitable causes with motions in Rotterdam

	Charitable cause	Alignment predicted by: Authors	Alignment predicted by: Debaters
	During periods of national housing shortages, this House would forcibly take ownership of privately owned homes which are not lived in by their owners)		
Motion 1	**Action on Empty Homes** is a UK NGO campaigning for more empty homes to be brought into use for people in housing need. It raises awareness of the waste of long-term empty homes and campaigns for changes to national policy to bring more homes into use.	proposition	Opposition (52%)
Motion 1	**The Land Is Ours** campaigns peacefully for access to the land, its resources, and the decision-making processes affecting them. Among other things, it advocates 'Use It Or Lose It' programme where empty buildings are forfeit or put on a tax escalator, where the owner can lose title after one year	proposition	Opposition (37%)
	This House believes that states should aggressively fund geoengineering projects instead of attempting to mitigate the effect of climate change		
Motion 2	**Geoengineering Monitor** aims to be a timely source for information and critical perspectives on climate engineering. The goal is to serve as a resource for people around the world who are opposing climate geoengineering and fighting to address the root causes of climate change instead	Opposition	proposition (54%)
Motion 2	**The Environmental Defense Fund** addresses today's most urgent environmental challenges by focusing on the solutions that will have the biggest impact, such as removing obsolete rules that hamper the clean energy market in the U.S. It favors a strategy of reducing CO2 emissions over geoengineering	Opposition	proposition (74%)
	This House regrets the decision to let the FARC (i.e The Revolutionary Armed Forces of Colombia - People's Army) run as a political party.		
Motion 3	**Justice for Colombia** is a British NGO whose primary goal is to give a political voice internationally to Colombian civil society. It has been campaigning to help Jesús Santrich, a lead FARC negotiator of the peace deal who was going to take a seat into parliament in 2018, get justice. The US incarcerated him without providing any evidence of Santrich's crime to the Colombian government	Opposition	proposition (69%)
Motion 3	**Strangers to Peace** is a documentary project of film maker Noah DeBonis which follows the life of ex-FARC guerrillas during their reintegration process. If funded, the film aims to enrich viewer's understanding of a marginalized community through tales of personal and social redemption	Opposition	proposition (63%)
	When tech companies own platform utilities and platform products, this House would break them up.		
Motion 4	**Elizabeth Warren** is a candidate for the President of the United States in 2020. Among other causes, she runs on a platform breaking up big tech firms such as Google and Amazon in a platform component and a supplier component. Donations go towards her campaign for the presidency	proposition	proposition (46%)
Motion 4	**The Open Markets Institute** uses journalism to promote greater awareness of the political and economic dangers of monopolization, identifies the changes in policy and law that cleared the way for such consolidation, and fosters discussions with policymakers and citizens as to how to update America's traditional political economic principles for our 21st century digital society	proposition	Opposition (42%)

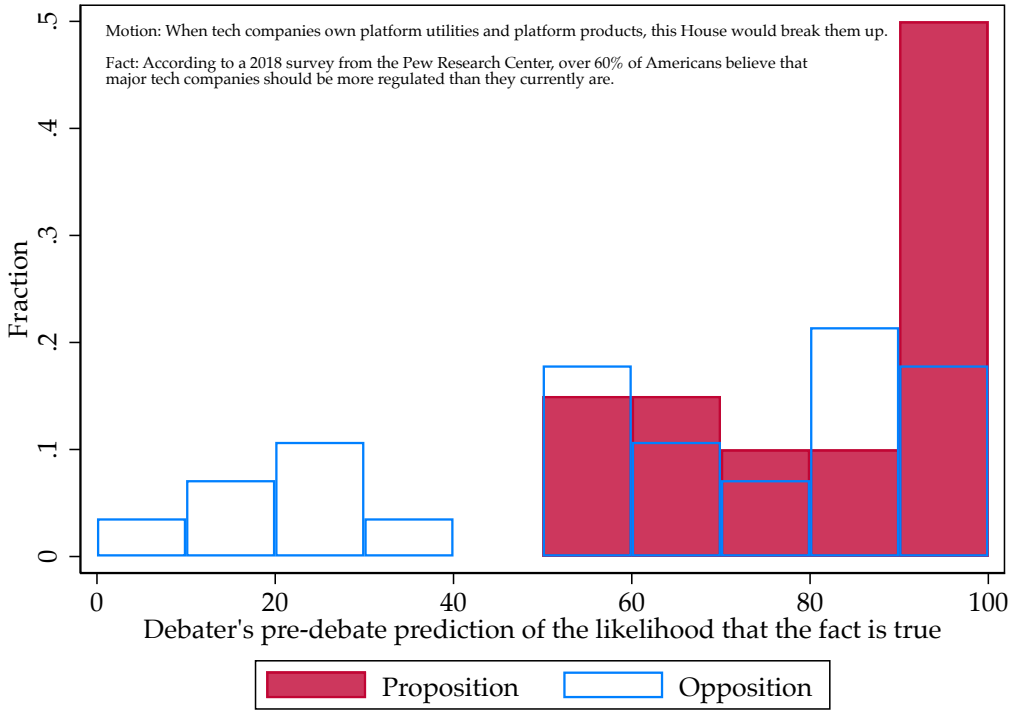
A.11 Variable Transformations

A.11.1 Beliefs regarding topics of the motions

The beliefs that we elicit for facts that are relevant to a motion are expected to capture alignment with either side of the motion. While in some cases we expect that someone who is aligned with the proposition is more likely to believe that a fact is true, in some other cases alignment with proposition is expected to be associated with a belief that a fact is false. Figure A.10 illustrates the example of a fact that we were expecting to capture alignment with the proposition. To half of the debaters in Rotterdam we asked this question just before the debate (predebate), and to another half after the debate. As the figure illustrates, in the predebate survey proposition speakers are more likely than Opposition speakers to believe that a survey conducted by the Pew Research Center in 2018 found that over 60% of Americans want major tech companies to be more regulated. The motion of this debate was that "When tech companies own platform utilities and platform products, this House would break them up.".

In order to make belief elicitations comparable across motions, we conduct a normal standardization of the reported belief (separately for each factual question asked at each survey), and we adjust the sign of the standardized belief in such a way that a positive (negative) sign of the standardized outcome captures alignment with the proposition (opposition) side of the motion. While we had a strong prior on the direction of alignment that each fact would capture, to make this sign correction objective and transparent we use the modal alignment predicted by debaters in the endline survey. Our predicted alignment and debaters' are reported in Table A.17 and Table A.20.

Figure A.10: Example of Reported Predebate Beliefs, by Side of the Debate



A.11.2 Attitudes regarding topics of the motions

Attitudes towards the motion are measured through an allocation of donations that individual debaters can make between a neutral charity – a charity that is used for every motion with an agenda that is relatively orthogonal to alignment with the motion, and a motion charity – a charity that is specific to each motion with an agenda that is expected to be particularly valued by an individual who is aligned with a particular side of the motion.

We had planned to follow a similar procedure as for beliefs to harmonize attitudes across motions. We diverge from that plan for two reasons: First, possible charitable allocations follow a discrete distribution, which clearly strongly violates normality. Second, due to poor phrasing of the mapping alignment question, answers to this question were very noisy and often conflicted with our prediction of alignment of the charity to the motion in ways that are hard to rationalize. In Table A.18 and Table A.21 we list for each charitable cause our predicted alignment with the motion as well as the debaters'.

Figure A.11: Example of Charity Allocations Chosen Predebate, by Side of the Debate

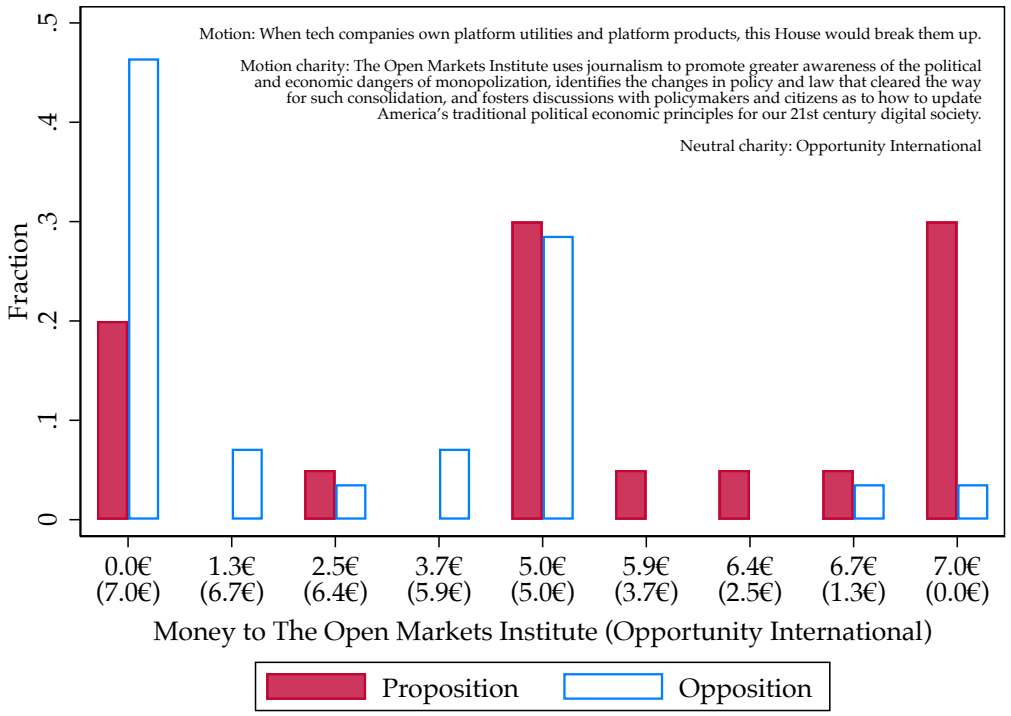


Figure A.11 illustrates an instance in which our prediction of alignment of the charity contradicts the debaters' as captured by the mapping question at endline: We predicted alignment of the motion charity with the proposition, while debaters predicted alignment of the motion charity with the Opposition. In this instance, debaters can choose to allocate money between a neutral charity, Opportunity International, and The Open Markets Institute, an NGO promoting awareness on the dangers of monopolization in the tech sector. From behavioral outcomes elicited predebate, we find that debaters tend to give more to The Open Markets Institute when they *propose* a motion that would break up big tech companies: the alignment that we predicted.

We decide to construct an harmonized ordinal variable that captures alignment with the proposition side of the motion using our predicted alignments. Such variable, for each question, simply takes the nine categories of increasing monetary amounts that are given to the baseline charity (and subtracted to the motion charity), and adjusts the order in such a way that if the motion charity is aligned with the proposition

(opposition) the order is reversed (kept as it is).

A.12 Mapping Pre-Analysis Plan to Paper

This study was pre-registered the week prior to the first debating competition. Relative to the pre-registered sample size and survey items we report the following substantial changes:

- We expected to have 104 teams of debaters across the two tournaments. We end up with 4 teams fewer in Rotterdam due to last minute cancellations.
- Dropped debaters' attractiveness score from the enumerator survey.

Pre-registration included a pre-analysis plan. In this appendix we spell out the analysis planned and the results of the planned analyses, which are sometimes replaced in the main paper with analyses that are now considered superior by the authors for statistical and expositional reasons.

A.12.1 Pre-registered Hypotheses

We formulated a first set of hypotheses deriving from strategic self-deception, and a second set of hypotheses on the role of debating for belief convergence.

Self-Persuasion

Hypothesis 4. *Debaters predebate factual beliefs are biased in the direction of their persuasion goal.*

The pre-registration specifies how beliefs are standardized and sign-adjusted to obtain a metric $b_{i,m}$ and conduct a fixed effects panel analysis to identify the causal effect of persuasion goals. Sign adjustment is determined by Endline responses to mapping questions in which, for each factual question and charity related to the motion, we ask subgroups of debaters to predict what the majority of respondents would believe the alignment to be between proposition/opposition/No alignment. When at least 51 percent of debaters correctly predict the reported modal alignment, we use that alignment

to determine the sign adjustment of standardized beliefs.[12] We test the hypothesis by estimating the following fixed effects model

$$b_{i,m} = \alpha_i + \beta proposition_{i,m} + \delta_m + \epsilon_{i,m}$$

in which δ_m are motion fixed effects, and $\epsilon_{i,m}$ is the error term allowing for a team component. Column (1) of Table 1.3 reports the estimated β from such model that confirms the original hypothesis, along with multiple additional specifications to assess the robustness of the result.

Hypothesis 5. *Debaters predebate attitudes are biased in the direction of their persuasion goal.*

The pre-registration specifies a similar standardization and sign-adjustment for our measure of attitudes, and a similar analysis of the causal effect of persuasion goals. Here we need to deviate from the pre-analysis plan. First, in the pre-analysis plan we failed to account for the ordinal nature of our attitudinal outcome, which does not warrant standard normalization. Therefore, we decide to conduct sign-adjustment, but not standardization. Second, we failed to adequately formulate the endline alignment question for charities. This led to puzzling alignment predictions presented in Table A.18 and Table A.21, that often conflict with our own prediction of alignment. Therefore, we decide to use the prediction of alignment formulated by us – that guided the choice of motion related charities in the first place. We test the hypothesis by estimating the following random effects model for the latent variable underlying our sign adjusted attitudinal outcome $a_{i,m}$:

$$\tilde{a}_{i,m} = \alpha_i + \beta proposition_{i,m} + \gamma X_i + \delta_m + \epsilon_{i,m}$$

in which X_i includes all socio-demographic and experience controls, δ_m are motion fixed effects, and $\epsilon_{i,m}$ is the error term allowing for a team component. Random effects models are used because standard fixed effects models for ordinal categorical vari-

[12] If the alignment of a belief distribution is proposition (opposition), then we change the sign of standardized beliefs for opposition (proposition) speakers.

ables are under-identified. Column (1) of Table A.7 reports the estimated β from such model without controls, column (2) reports estimates from the model with controls. Both estimates confirm the original hypothesis. We also report additional results from Chamberlain-like fixed effects estimators (column (3)) to assess the robustness of the result.

Hypothesis 6. *Debaters have more confidence in the arguments favoring their side than in the other side's arguments.*

The pre-registration specifies a straightforward fixed effects regression model to test this hypothesis using the prediction that the majority of debates in parallel debates will be won by proposition teams $c_{i,m}$:

$$c_{i,m} = \alpha_i + \beta proposition_{i,m} + \delta_m + \epsilon_{i,m}$$

in which δ_m are motion fixed effects, and $\epsilon_{i,m}$ is the error term allowing for a team component. Column (1) of Table 1.5 reports the estimated β from such model that confirms the original hypothesis, along with multiple additional specifications to assess the robustness of the result.

Hypothesis 7. *When persuasion goals are more aligned with private beliefs at baseline, debaters obtain higher persuasiveness ratings by judges.*

The pre-registration specifies a fixed effects regression model to test the correlation between *baseline alignment* and *persuasiveness*, where baseline alignment is defined as standardized and sign-adjusted baseline belief above 0 (below 0) if for speakers that will be assigned to proposition (opposition), and *persuasiveness* as the panel average of the independent scores that each judges gives for broad persuasiveness of speaker's performance $P_{i,m}$:

$$P_{i,m} = \alpha_i + \beta(\mathbb{1}_{y_{i,m}^{baseline} \geq 0} \mathbb{1}_{proposition_{i,m}} + \mathbb{1}_{b_{i,m}^{baseline} < 0} \mathbb{1}_{Opposition_{i,m}}) + \delta_m + \epsilon_{i,m}$$

in which δ_m are motion fixed effects, and $\epsilon_{i,m}$ is the error term allowing for a team

component. Column (1) of Table A.10 reports the estimated β from such model that lends no support for such hypothesis.

Debating and Convergence

Hypothesis 8. *Postdebate attitudes are less dispersed than predebate attitudes.*

The pre-registered analysis proposes to assess whether an individual level measure distance from the median ordinality of sign-adjusted bundle $d(a)_{i,m,p,s}$ is lower at postdebate than it is predebate.[13] We test for convergence of attitudes in the following fixed effects regression framework:

$$d(a)_{i,m,p,s} = \alpha_i + \beta Predebate_{i,m,p} + \delta_p + \delta_m + \epsilon_{i,m,p,s}$$

in which δ_m are motion fixed effects, δ_p are charity-pair fixed effects, and $\epsilon_{i,m,p,s}$ is the error term allowing for a team component. We would say that there is convergence in attitudes from predebate to postdebate if β were positive and significant. Column (9) of Table A.3 reports the estimated β from such model that lends no statistically significant support for such hypothesis.

Hypothesis 9. *Postdebate factual beliefs are less dispersed than predebate and baseline beliefs.*

The pre-registered analysis proposes to assess whether an individual level measure distance from the median ordinality of sign-adjusted bundle $d(b)_{i,m,q,s}$ is lower at postdebate than it is at predebate and baseline.[14] We test for convergence of beliefs from predebate to postdebate in the following fixed effects regression framework:

$$d(b)_{i,m,q,s_1} = \alpha_i + \beta_1 Predebate_{i,m,p} + \delta_p + \delta_m + \epsilon_{i,m,p,s_1}$$

and for convergence of beliefs from baseline to postdebate in the following fixed effects

[13] For a sign-adjusted distribution of monetary donations to charitable organizations taking place at survey s of motion m for pair of charities p, $d(a)_{i,m,p,s} = |a_{i,m,p,s} - median(a_{i,m,p,s})|$.

[14] For a distribution of beliefs elicited at survey s of motion m for factual question q, $d(b)_{i,m,q,s} = |b_{i,m,q,s} - median(b_{i,m,q,s})|$.

regression framework:

$$d(b)_{i,m,q,s_2} = \alpha_i + \beta_2 Predebate_{i,m,p} + \delta_p + \delta_m + \epsilon_{i,m,p,s_2}$$

in which $s_1 \in \{Predebate, Postdebate\}$, $s_2 \in \{Baseline, Postdebate\}$, δ_m are motion fixed effects, δ_p are charity-pair fixed effects, and $\epsilon_{i,m,p,s}$ is the error term allowing for a team component. We would say that there is convergence in attitudes from Predebate (Baseline) to Postdebate if β_1 (β_2) were positive and significant. Column (3) and (5) of Table A.3 report the estimated β_1 and β_2 from such models, respectively. The estimate of β_1 rejects the null hypothesis of convergence in a one-sided t-test, and provides evidence that beliefs in fact polarize from Baseline to Postdebate. The estimate of β_2 is qualitatively in line with convergence, but not statistically different from zero.

Hypothesis 10. *Postdebate factual beliefs are less dispersed than predebate and baseline beliefs, looking at only those debaters who got to argue their baseline position.*

The plan for testing this hypothesis was to exactly replicate the analysis for Hypothesis 9, including in the analysis only the distance in beliefs from the median belief for debaters that have at baseline standardized and sign-adjusted beliefs aligned with their persuasion goal. Column (6) and (7) of Table A.3 report the estimated β_1 and β_2 from the estimates of the regression models for such sub-sample, respectively. The estimate of β_1 rejects the null hypothesis of convergence in a one-sided t-test, and provides evidence that beliefs in fact polarize from Baseline to Postdebate. The estimate of β_2 is qualitatively in line with convergence, but not statistically different from zero.

Hypothesis 11. *Heated debates are less likely to favor the formation of a consensus around facts and attitudes, and may even increase polarization.*

The plan for testing this hypothesis was to exactly replicate the analysis for Hypothesis 8 and Hypothesis 9, including in regression analysis an interaction term between the timing of the elicitation (the survey dummy) and a binary indicator for whether a debater was heated or not.

Appendix B

Appendix to Chapter 2

B.1 Theoretical Appendix

B.1.1 More general framework

Consider the following more general formulation of the model presented in section 2.2.2 for any distribution of types $F(v)$ and cost function $c(\cdot)$ such that $c'(\cdot) > 0$, $c''(\cdot) > 0$ and $c(0) = 0$.

$$U(d_i) = (v_i + m_i)d_i - c(d_i) - \frac{\lambda_{i,j}}{2} Pr(\mathcal{A}_j(m_j))(d_i - E(d_j^n|\mathcal{A}_j(m_j)))^2$$

where $\mathcal{A}_j(m_j) = \{v_j \in V : c(d_j^n)/d_j^n > m_j\}$. Assuming that $Pr(\mathcal{A}_j(m_j))$ and $E(d_j^n|\mathcal{A}_j(m_j))$ are continuous in m_j, the model gives the following comparative statics for the effect of peer's incentives on individual donations:

$$\frac{\partial d_i^*}{\partial m_j} = \Lambda \left[\frac{\partial Pr(\mathcal{A}_j(m_j))}{\partial m_j} \left(E(d_j^n|\mathcal{A}_j(m_j)) - d_i^* \right) + Pr(\mathcal{A}_j(m_j)) \frac{\partial E(d_j^n|\mathcal{A}_j(m_j))}{\partial m_j} \right]$$

where $\Lambda = \frac{\lambda}{c''(d_i^*) + \lambda Pr(\mathcal{A}_j(m_j))}$ and $\mathcal{A}_j(m_j) = \{v_j \in V : c(d_j^n)/d_j^n > m_j\}$. This result leads to the following Lemma:

Lemma 1. *For cost functions with constant elasticity of donation effort $k \leq 1$, (i) peer's incentives tend to increase (decrease) individual donations for the more (less) altruistic and those with higher (lower) private incentives m_i to donate. However, (ii) there exists a $\tilde{m}_j$ threshold above which $\frac{\partial d_i^*}{\partial m_j}$ is 0 for any i.*

Part (i) of the lemma follows from the observation that a necessary condition for $\frac{\partial d_i^*}{\partial m_j}$ to be negative when $\frac{\partial Pr(\mathcal{A}_j(m_j))}{\partial m_j}$ is negative is that $E(d_j^n|\mathcal{A}_j(m_j)) \geq d_i^*$. This prediction is specific to this model and runs counter to models of (impure) altruism where donations are more likely to be strategic substitutes for more altruistic donors. To see

why this happens, notice that conformity compels less altruistic individuals to donate more than they would like, and increasing incentives for the social reference has two effects on the utility loss from not conforming $Pr(\mathcal{A}_j(m_j))(d_i - E(d_j^n|\mathcal{A}_j(m_j)))^2$: First, an increase due to the increase in d_j^n; second, a decrease due to the decrease in $Pr(\mathcal{A}_j(m_j))$. The latter can be seen as alleviating the pressure to conform and helps individuals adjust their donations towards their natural type.

Part (ii) of the Lemma requires proof, which we provide below.

Proof. The main step of this proof is to show that when the elasticity of donation effort with respect to the private value of effort $v_j + m_j$ is less or equal to 1, $Pr(\mathcal{A}_j(m_j))$ is decreasing and $E(d_j^n|\mathcal{A}_j(m_j))$ is increasing in m_j.

Because $\frac{c(d_j^n)}{d_j^n}$ is monotonically increasing in m_j and v_j, all we need for $\frac{\partial Pr(\mathcal{A}_j(m_j))}{\partial m_j} \leq 0$ is to show that the marginal altruistic types are excluded from $\mathcal{A}_j(m_j)$ when m_j increases.

That is, take $v_j^{'}$ and $m_j^{'}$ s.t.

$$\frac{c(d_j^n(v_j^{'}, m_j^{'}))}{d_j^n(v_j^{'}, m_j^{'})} = m_j^{'},$$

we want to show that a positive h implies

$$\frac{c(d_j^n(v_j^{'}, m_j^{'} + h))}{d_j^n(v_j^{'}, m_j^{'} + h)} \leq m_j^{'} + h.$$

Notice that this is always the case if $\frac{\partial}{\partial m_j}\left[\frac{c(d_j^n)}{d_j^n}\right] \leq 1$. With that in mind, consider

$$\frac{\partial}{\partial m_j}\left[\frac{c(d_j^n)}{d_j^n}\right] = \frac{(c'(d_j^n)d_j^n - c(d_j^n))\frac{\partial d_j^n}{\partial m_j}}{(d_j^n)^2} = \frac{(v_j + m_j)}{d_j^n}\frac{\partial d_j^n}{\partial m_j} - \frac{c(d_j^n)}{d_j^n}\frac{\frac{\partial d_j^n}{\partial m_j}}{d_j^n}$$

which for cost functions characterized by constant elasticity k of donations with re-

spect to the the value of donations $(v_j + m_j)$ can be rewritten as[1]

$$\frac{\partial}{\partial m_j}\left[\frac{c(d_j^n)}{d_j^n}\right] = k - \frac{c(d_j^n)}{d_j^n}\frac{k}{(v_j + m_j)}$$

using $c''(d) > 0$, $c'(d) > 0$, and $c(0) = 0$, we know that $0 \leq \frac{c(d_j^n)}{d_j^n} \leq (v_j + m_j)$. In turn, $k \leq 1$ is sufficient condition for $\frac{\partial}{\partial m_j}\left[\frac{c(d_j^n)}{d_j^n}\right] \leq 1$.

Under the same conditions, $E(d_j^n|\mathcal{A}_j)$ is increasing in m_j. This is because, as we have shown, larger incentives drive less altruistic types out of $\mathcal{A}_j(m_j)$ and because $\frac{\partial d_j^n}{\partial m_j} > 0$ for any v_j. □

B.1.2 Impure Altruism

In this section, we consider the properties of a model in which individuals have decreasing marginal utility from aggregate donations to the charity. We model this as

$$u_i = (v_i + m_i)d_i + g(d_i + d_j) - \frac{c}{2}d_i^2 \tag{B.1}$$

where we can now distinguish between warm-glow and pure altruism in an impure altruism function that is linear in warm glow $v_i d_i$ and has pure altruism $g(D)$ with $g'(D) > 0$, but $g''(D) < 0$. In addition, we require that $|\frac{cg''(D)}{c-g''(D)}| < 1$ in order to guarantee interior solutions. The Nash equilibrium in this game is characterized by an analogue to equation 2.3:

$$d_i = \frac{m_i + v_i + g'(d_i + d_j)}{c} \tag{B.2}$$

and the corresponding condition for j. Performing comparative statics on B.2 show that an agent's donations are increasing in her own incentives, holding those of her

[1] Using

$$k = \frac{v_j + m_j}{d_j^n}\frac{\partial d_j^n}{\partial(v_j + m_j)} = \frac{v_j + m_j}{d_j^n}\frac{\partial d_j^n}{\partial m_j}$$

peer constant:

$$\frac{\partial d_j}{\partial m_j} = c\frac{c - g''}{(c - g'')^2 - (cg'')^2} > 0 \tag{B.3}$$

Notice that the denominator in B.3 is positive because of the existence condition for interior solutions. For the impact of the peer's incentive on an agent, we find that

$$\frac{\partial d_i}{\partial m_j} = \frac{cg''}{c - g''}\frac{\partial d_j}{\partial m_j} < 0 \tag{B.4}$$

Thus, with (global) diminishing marginal utility from donations to the charity, an agent's donations are strictly decreasing in her peer's incentives, as claimed in the text. □

B.1.3 Incentive Inequality

One possible objection to leveraging heterogeneous monetary incentives to act prosocially for investigating the conformity channel of social influence is that incentive inequality in itself could be a source of strategic complementarities in donations. Recent research from Breza, Kaur, and Shamdasani, 2017 shows that unjustifiably heterogeneous incentives in work environment can introduce a form of inequity aversion (Fehr and Schmidt, 1999) that damages morale to exert effort. The morale effect of incentive inequality is a salient form of inequity aversion even when opportunities for comparing realized payoffs are limited, and remains meaningful when payoff disparities depend on effort (rather than allocation decisions). In this section, we illustrate when this form of inequity aversion can induce strategic complementarities in donations.

Consider a simple model of prosocial behavior similar to the one presented in section B.1.1, and replace the conformity term with the morale utility term from Breza, Kaur, and Shamdasani, 2017.

$$U(d_i) = (v_i + m_i)d_i - \frac{c}{2}d_i^2 + M(m_i, m_j)d_i \tag{B.5}$$

Morale $M(\cdot)$, as illustrated below, is a function of the gap in incentives between i and j, and allows for additional direct psychological incentive effects. Parameters α and β capture the extent to which people differentially dislike disadvantageous and advantageous inequality, respectively. The function $g(m_i)$ captures any sort of direct psychological effects of incentives, and $f(\cdot)$ is monotonically increasing in the gap between incentives and satisfies $f(0) = 0$.

$$M(m_i, m_j) = g(m_i) - \alpha f(m_i - m_j | m_i < m_j) - \beta f(m_j - m_i | m_i > m_j)$$

From this simple model we can derive the closed form of the optimal donation level, which is interpreted in the prediction that follows.

$$d_i^* = c^{-1} \left[v_i + m_i - \alpha f(m_i - m_j | m_i < m_j) - \beta f(m_j - m_i | m_i > m_j) + g(m_i) \right]$$

Prediction 1. *[Incentive Inequality] If donor's morale is damaged by incentive inequality, (i) at any m_i, i's donations are monotonically decreasing in the size of incentive inequality, and (ii) an increase (decrease) in either i's or j's incentives that reduces (increases) incentive inequality increases (decreases) donations of both i and j.*

The obvious implication of (i) is what we label a *main diagonal condition*: holding i's incentives constant, i's donations should be highest when incentives are homogeneous, and monotonically decreasing in the size of the $m_i - m_j$ gap.

Part (ii) further illustrates when incentive inequality introduces strategic complementarities in donations. However, notice how an increase (decrease) in m_j that accentuates (reduces) the gap between m_i and m_j decreases (increases) d_i and has a mixed effect on d_j – strengthening the strategic substitution of donations when the direct incentive effect on d_j dominates the negative (positive) effect of increased (decreased) inequality on j's morale.

B.2 Empirical Appendix

B.2.1 Morale Effects of Incentive Inequality

In this section, we test for the morale effects of incentive inequality using a joint test of the *main diagonal condition* that the model in Section B.1.3 implies.

The test of this joint hypothesis builds on Burks et al., 2009. We treat average donations in the nine incentivized treatments of our experiment as a nine-dimensional normal distribution with means $\boldsymbol{\mu}_{p_i,p_j}$ (which we treat as unknown) and diagonal covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\sigma^2}_{p_i,p_j}\mathbb{I}$ (which we treat as known). For the joint test, we use maximum likelihood to determine the vector $\hat{\boldsymbol{\mu}}_{p_i,p_j}$ that best fits the nine dimensional vector of sample means $\overline{\boldsymbol{Donation}}_{p_i,p_j}$ - with and without the inequality constraints imposed by the *main diagonal condition*. A Likelihood Ratio test from the constrained and unconstrained likelihood functions is used to jointly assess these constraints. The test statistic is $\chi^2_{(d)}$ distributed with degrees of freedom d equal to the number of binding inequality constraints.

Table B.2.1 reports the raw first moments of the nine-dimensional distribution, the moments estimated with constrained Maximum Likelihood (constrained by the main diagonal condition), and the corresponding Likelihood Ratio tests. Both for the whole sample, and splitting the sample by oneness. Looking at the whole sample, one can notice qualitative violations of the main diagonal condition that cause the constrained estimates of the first moments to differ from the raw means. However, such violations are not sufficiently strong to reject the joint hypothesis ($p = 0.391$).

Next, we test the restriction on the samples split by oneness. In panel (b), we confirm that the restrictions imposed by inequity aversion cannot be rejected among low oneness subjects ($p = 0.737$). In panel (c), we strongly reject the *main diagonal condition* among high oneness subjects ($p = 0.002$). To understand how inequity aversion is rejected for more socially close subjects, it is worth interpreting the two main local violations that determine the results of the joint test. The first local violation is due to the change in average donations between groups of players who get randomized out

of incentives: increases in their peer's incentives – that *ceteris paribus* increase incentive inequality – increase their own donations. This result is clearly inconsistent with the morale effects of incentive inequality, and is also inconsistent with a concave altruistic utility of giving.[2] The second local violation is due to the change in average donations between groups of players who get randomized into relatively high incentives (*good news*): decreases in their peers' incentives – that *ceteris paribus* increase incentive inequality – increase their own donations. This result is significant for the decrease in peers' incentives from high to moderate, and may be explained by substitution due to concave (altruistic) utility of giving. However, evidence that expectations about peers' levels of giving are virtually identical between these two groups makes this explanation unlikely.

[2]The standard framework of inequity aversion (Fehr and Schmidt, 1999), is less tractable in our setting because realized payoff inequality depends both on incentives provided and effort choices. Such a framework does however make the clear prediction that peer's incentives should not affect individual donations when an agent gets no incentives, and the t-test for one of the two local violations ($\hat{\mu}_{n,n} = \hat{\mu}_{n,m}$) reported in Table B.2.1 panel (c) provides evidence against this prediction.

Table B.2.1: Average Donations in Lottery Treatments, Maximum Likelihood Estimates (Coefficient Estimates and Standard Errors in Parentheses)

(a) Full sample

Incentives to self	Data: Incentives to peer Zero (1)	Data: Incentives to peer Moderate (2)	Data: Incentives to peer High (3)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Zero (4)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Moderate (5)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer High (6)	Main Diagonal	p-value (7)
Zero	3.233	3.417	3.190	3.320	3.320	3.190	LR: $\chi^2_{(2)} = 1.877$	0.391
	(0.217)	(0.230)	(0.209)	(0.217)	(0.230)	(0.209)		
Moderate	5.042	5.546	5.155	5.042	5.546	5.155	Local Violations: t-tests	
	(0.233)	(0.235)	(0.224)	(0.233)	(0.235)	(0.224)	H0: $\hat{\mu}_{n,n} = \hat{\mu}_{n,m}$	0.551
High	5.299	5.575	5.187	5.299	5.366	5.366	H0: $\hat{\mu}_{h,m} = \hat{\mu}_{h,h}$	0.215
	(0.233)	(0.229)	(0.212)	(0.233)	(0.229)	(0.212)		

(b) Low oneness

Incentives to self	Data: Incentives to peer Zero (1)	Data: Incentives to peer Moderate (2)	Data: Incentives to peer High (3)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Zero (4)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Moderate (5)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer High (6)	Main Diagonal	p-value (7)
Zero	3.190	2.667	2.593	3.190	2.667	2.593	$\chi^2_{(1)} = 0.113$	0.737
	(0.320)	(0.304)	(0.299)	(0.320)	(0.304)	(0.299)		
Moderate	4.778	5.105	4.622	4.778	5.105	4.622	Local Violations: t-tests	
	(0.337)	(0.339)	(0.332)	(0.337)	(0.339)	(0.332)	H0: $\hat{\mu}_{h,n} = \hat{\mu}_{h,h}$	0.727
High	5.549	4.889	5.382	5.456	4.889	5.456		
	(0.370)	(0.323)	(0.331)	(0.370)	(0.323)	(0.331)		

(c) High oneness

Incentives to self	Data: Incentives to peer Zero (1)	Data: Incentives to peer Moderate (2)	Data: Incentives to peer High (3)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Zero (4)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer Moderate (5)	$\hat{\theta}^{ML}_{constrained}$: Incentives to peer High (6)	Main Diagonal	p-value (7)
Zero	3.270	4.099	3.614	3.635	3.635	3.614	$\chi^2_{(2)} = 12.443$	0.002
	(0.295)	(0.333)	(0.285)	(0.295)	(0.333)	(0.285)		
Moderate	5.263	5.913	5.627	5.263	5.913	5.627	Local Violations: t-tests	
	(0.322)	(0.323)	(0.298)	(0.322)	(0.323)	(0.298)	H0: $\hat{\mu}_{n,n} = \hat{\mu}_{n,m}$	0.057
High	5.103	6.293	5.034	5.103	5.581	5.581	H0: $\hat{\mu}_{h,m} = \hat{\mu}_{h,h}$	0.003
	(0.297)	(0.316)	(0.277)	(0.297)	(0.316)	(0.277)		

Notes: Degrees of freedom of the Likelihood Ratio test statistic equal the number of binding inequality constraints imposed by the composite null hypothesis. Empirical standard errors of the means are directly fed into the maximum likelihood routine.

The *main diagonal condition* has a mirror set of conditions on beliefs across treatments. Table B.2.2 also shows rejection of the conditions imposed by the morale effects on incentive in equality on beliefs.

Taken together, the results of the analyses highlight that the complementarities observed in the data are at variance with the predictions of inequity aversion. This contrast is particularly stark among subjects with closer connection to their peer, which leaves our conformity framework as the more plausible explanation for our findings.

Table B.2.2: Average Beliefs in Lottery Treatments, Maximum Likelihood Estimates (Coefficient Estimates and Standard Errors in Parentheses)

(a) Full sample		Data			$\hat{\theta}^{ML}_{constrained}$				
		Incentives to self			Incentives to self			Main Diagonal	
		Zero	Moderate	High	Zero	Moderate	High		p-value
		(1)	(2)	(3)	(4)	(5)	(6)		(7)
Incentives to peer	Zero	2.540	2.585	2.374	2.561	2.561	2.374	$\chi^2_{(2)} = 6.277$	0.043
		(0.182)	(0.193)	(0.174)	(0.182)	(0.193)	(0.174)		
	Moderate	4.331	4.832	4.100	4.331	4.832	4.100		
		(0.215)	(0.214)	(0.201)	(0.215)	(0.214)	(0.201)		
	High	4.637	5.086	4.374	4.637	4.708	4.708		
		(0.208)	(0.207)	(0.195)	(0.208)	(0.207)	(0.195)		

(a) Full sample		Data			$\hat{\theta}^{ML}_{constrained}$				
		Incentives to self			Incentives to self			Main Diagonal	
		Zero	Moderate	High	Zero	Moderate	High		p-value
		(1)	(2)	(3)	(4)	(5)	(6)		(7)
Incentives to peer	Zero	2.124	2.053	1.754	2.124	2.053	1.754	$\chi^2_{(1)} = 0.041$	0.840
		(0.241)	(0.264)	(0.225)	(0.241)	(0.264)	(0.225)		
	Moderate	3.703	3.992	3.357	3.703	3.992	3.357		
		(0.303)	(0.296)	(0.262)	(0.303)	(0.296)	(0.262)		
	High	3.907	4.301	4.213	3.907	4.256	4.256		
		(0.316)	(0.310)	(0.303)	(0.316)	(0.310)	(0.303)		

(a) Full sample		Data			$\hat{\theta}^{ML}_{constrained}$				
		Incentives to self			Incentives to self			Main Diagonal	
		Zero	Moderate	High	Zero	Moderate	High		p-value
		(1)	(2)	(3)	(4)	(5)	(6)		(7)
Incentives to peer	Zero	2.890	3.032	2.859	2.959	2.959	2.859	$\chi^2_{(3)} = 12.248$	0.007
		(0.265)	(0.272)	(0.249)	(0.265)	(0.272)	(0.249)		
	Moderate	4.901	5.530	4.878	4.901	5.530	4.878		
		(0.297)	(0.292)	(0.293)	(0.297)	(0.292)	(0.293)		
	High	5.157	5.783	4.500	5.124	5.124	5.124		
		(0.270)	(0.267)	(0.254)	(0.270)	(0.267)	(0.254)		

Notes: Degrees of freedom of the Likelihood Ratio test statistic equal the number of binding inequality constraints imposed by the composite null hypothesis. Empirical standard errors of the means are directly fed into the maximum likelihood routine.

B.2.2 Additional Tables

Table B.2.3: OLS for Determinants of Social Proximity (Coefficient Estimates and Standard Errors in Parentheses)

Outcome: Oneness scale	(1)	(2)
Contact		1.434***
		(0.057)
Male	0.120*	0.139**
	(0.062)	(0.056)
Same gender	0.236***	0.180***
	(0.061)	(0.056)
Age, absolute difference	-0.003	-0.001
	(0.003)	(0.003)
Experience, absolute difference	-0.080***	-0.072***
	(0.024)	(0.022)
Constant	3.051***	2.118***
	(0.122)	(0.116)
Observations	2914	2914
R^2	0.014	0.189
Correlation in regression residuals (oneness scale) between peers	0.294 (0.340)	0.167 (0.340)

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: All specifications include age group, experience, and session dummies. Standard errors are clustered at the pair level.

Table B.2.4: Incentive Effects on Donations

(Coefficient Estimates and Standard Errors in Parentheses)

Outcome: Donations	(1)	(2)
Lottery	-0.763***	-0.835***
	(0.288)	(0.282)
Incentives to self (baseline: *Zero*)		
Moderate	1.968***	1.968***
	(0.184)	(0.181)
High	2.090***	2.058***
	(0.180)	(0.177)
Incentives to peer (baseline: *Zero*)		
Moderate	-0.286	-0.195
	(0.258)	(0.255)
High	-0.287	-0.263
	(0.261)	(0.258)
Moderate × High oneness	1.170***	1.049***
	(0.349)	(0.343)
High × High oneness	0.486	0.445
	(0.332)	(0.328)
Constant	3.906***	4.680***
	(0.262)	(0.345)
Controls	No	Yes
Observations	2914	2914

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Specification with controls includes age group, experience, and session dummies. Standard errors are clustered at the pair level.

B.2.3 Additional Figures

Figure B.1: Joint Problem Solving Task Software Interface

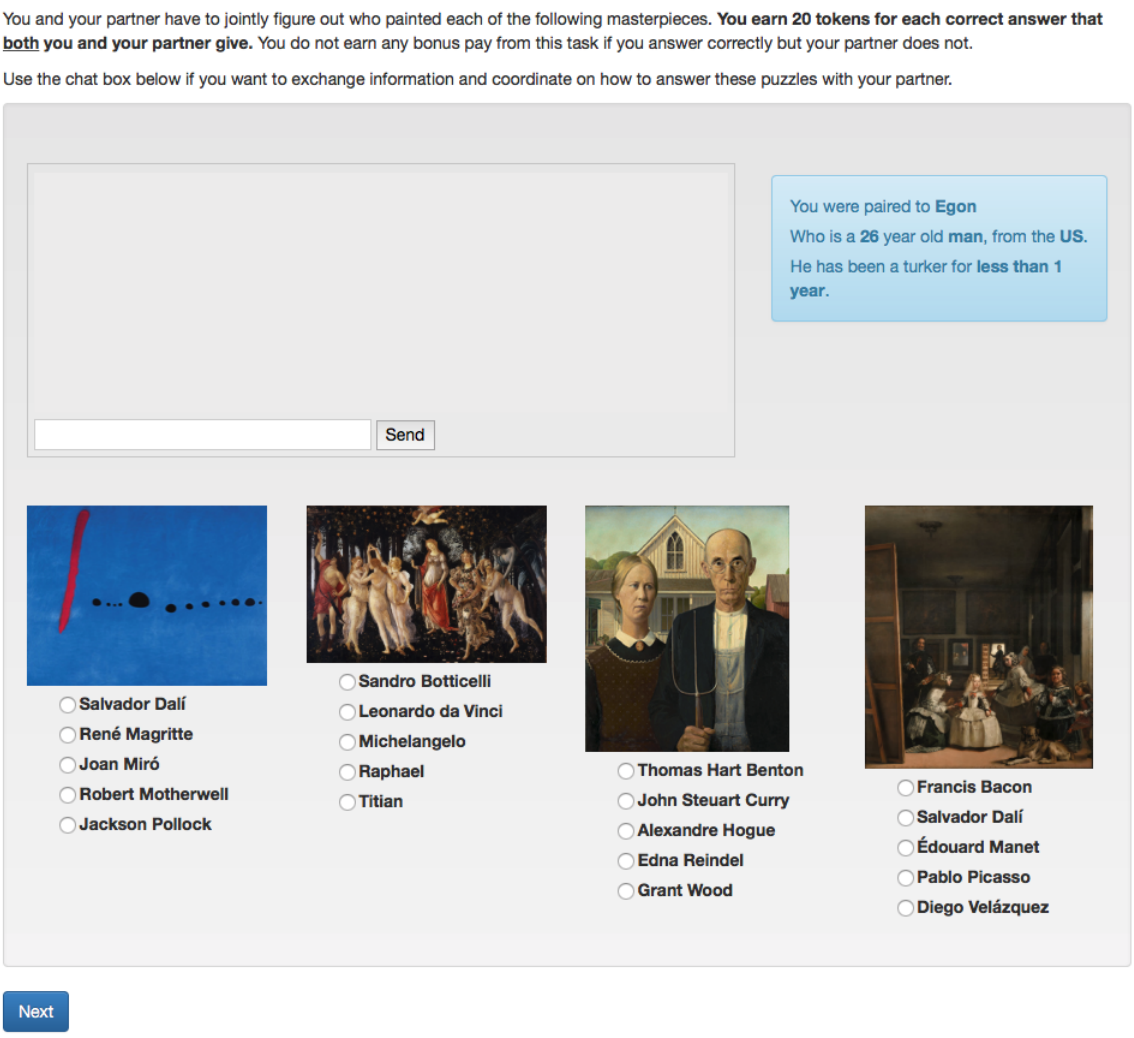


Figure B.2: Elicitation of the IOS (top) and WE (bottom) Scales

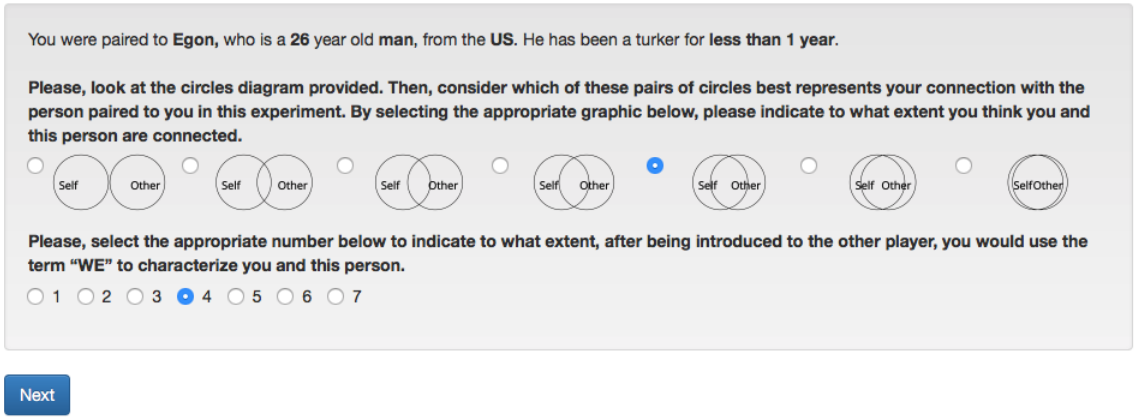


Figure B.3: Elicitation of Beliefs and Donations, and Treatment Assignment

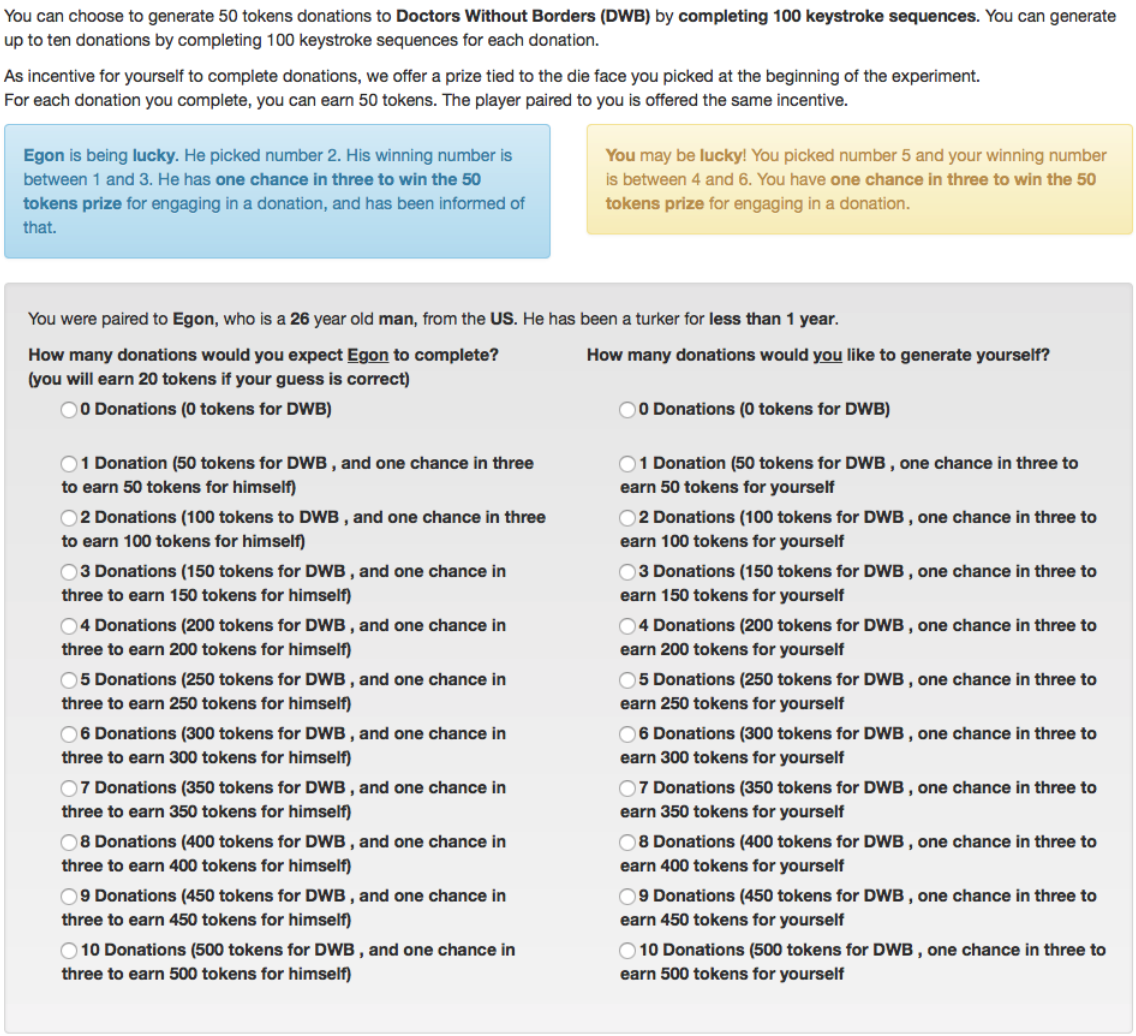


Figure B.4: Distribution of Social Proximity Scales

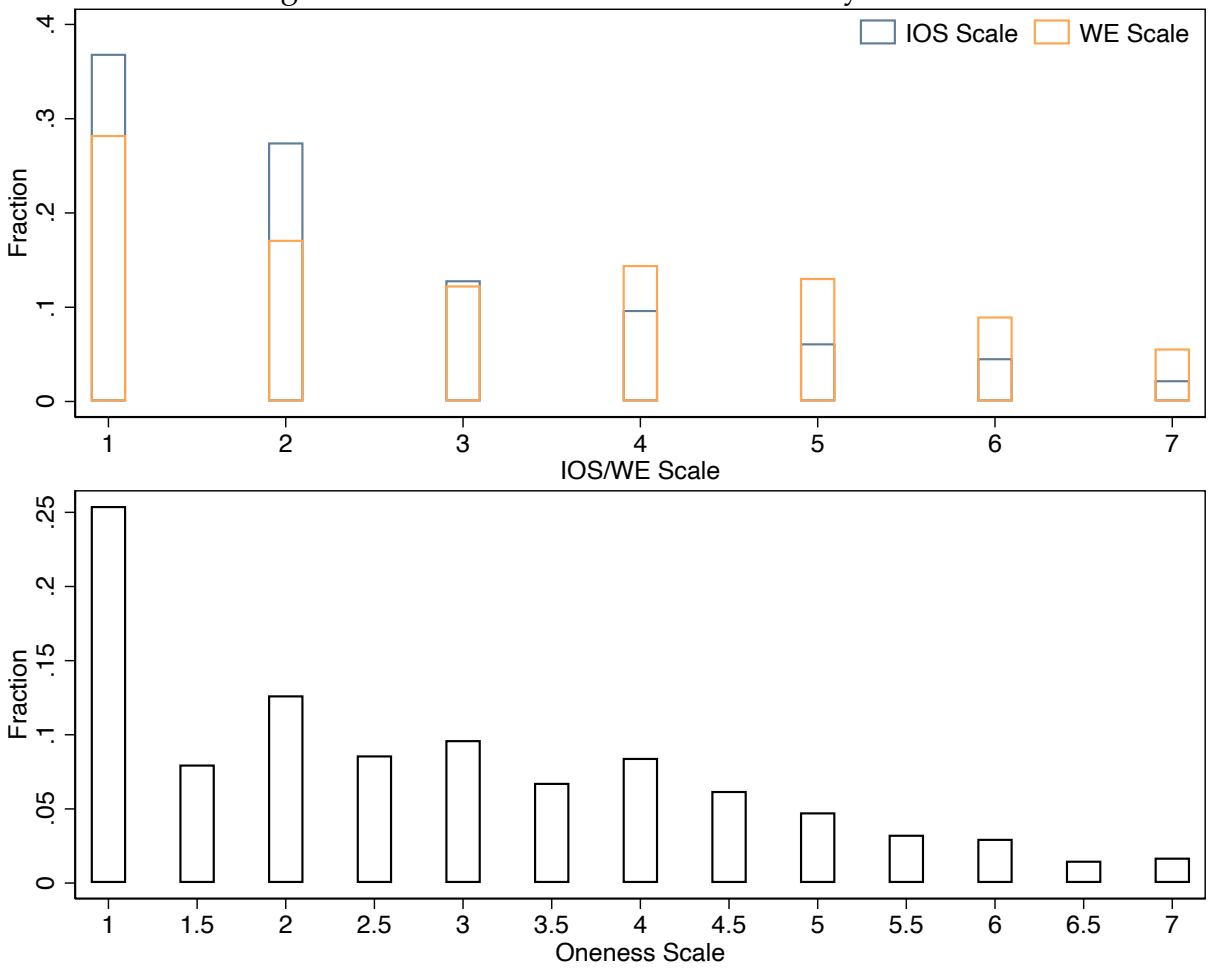
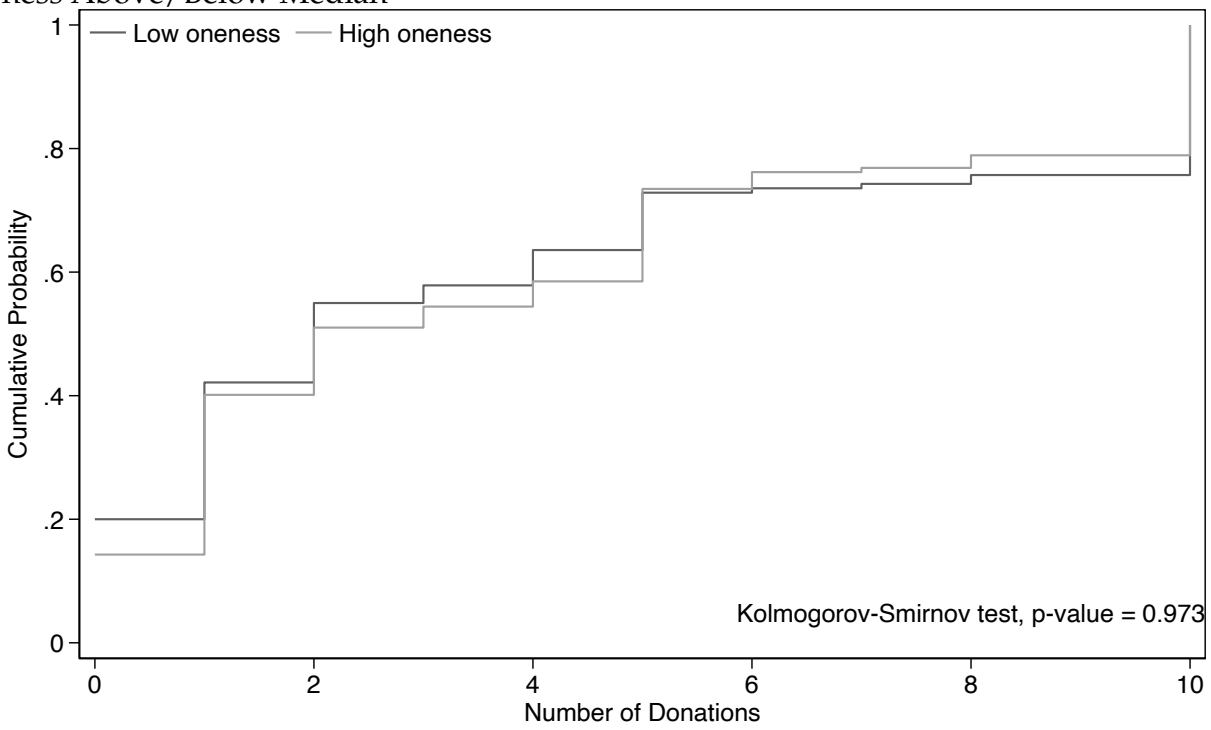


Figure B.5: Cumulative Density Function of Donations in Control Treatment, by Oneness Above/Below Median



B.3 Complete Instructions

B.3.1 Page 0: Consent

Please read this before clicking "Accept"

This HIT is an academic experiment on economic decision making. Based on how you play the experiment, we will donate money to a charitable organization.

By participating in this experiment, you are participating in a study performed by researchers at the University of Bonn. All data collected in this study are for research purposes only.

The experiment requires you to press keys on your keyboard. You thus need full dexterity in at least one hand. The experimental software complies with modern web standards, but may require a physical keyboard to detect your keystrokes. For part of the experiment you will be interacting with another player. To ensure that interactions occur in a timely manner we give each participant 5 minutes maximum to complete each of the following two pages. For the rest of the experiment a session timeout applies. Your session expires 40 minutes after you accept this HIT. If you do not want to complete the HIT within 40 minutes, we advise to return the HIT. We will not be able to approve work for timed out HITs.

Compensation: After completing this HIT, you will receive your reward plus a bonus payment that is based on how you play the experiment.

Legal information: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Any reports and presentations about the findings from this study will not include any information that could identify you. We may share the data we collect in this study with other researchers doing future studies; if we share your data, we will not include any information that could identify you. By accepting this HIT, you indicate that you are older than 18 years and agree to participate in this experiment.

B.3.2 Page 1: Introduction

In this study each participant will be given the opportunity to engage independently in a real effort game. By participating you create value for a charity.

Part of your variable bonus may be uncertain. For this we will ask you to pick a number between 1 and 6, which the experimental software will match to a digital roll of die.

What face of a die would you pick? [drop-down list]

For this experiment you will be paired to another player that is currently participating in the same experiment. Given that part of the experiment will involve common problem solving, we would like pairs of players to get to know each other. For this, on the next page we are collecting some basic socio-demographic information, which will be shared with the paired participant. The socio-demographic information collected is minimal and does not make you personally identifiable.

Throughout the experiment, you will engage in tasks that will determine your variable bonus. Completing tasks you accumulate tokens. Tokens are converted to USD at the end of the HIT. One token is worth 0.005 USD.

This experiment is a research effort to understand economic behavior. In what follows there will be no deception: we will do nothing different from what is explained to you. For any question do not esitate to contact us.

B.3.3 Page 2: Survery on Demographic Information

We would like paired players to know a bit about each other. For this, we are collecting some basic socio-demographic information, which will be shared with the other participant.

What is your first name? [text field]

What is your age? [drop-down list]

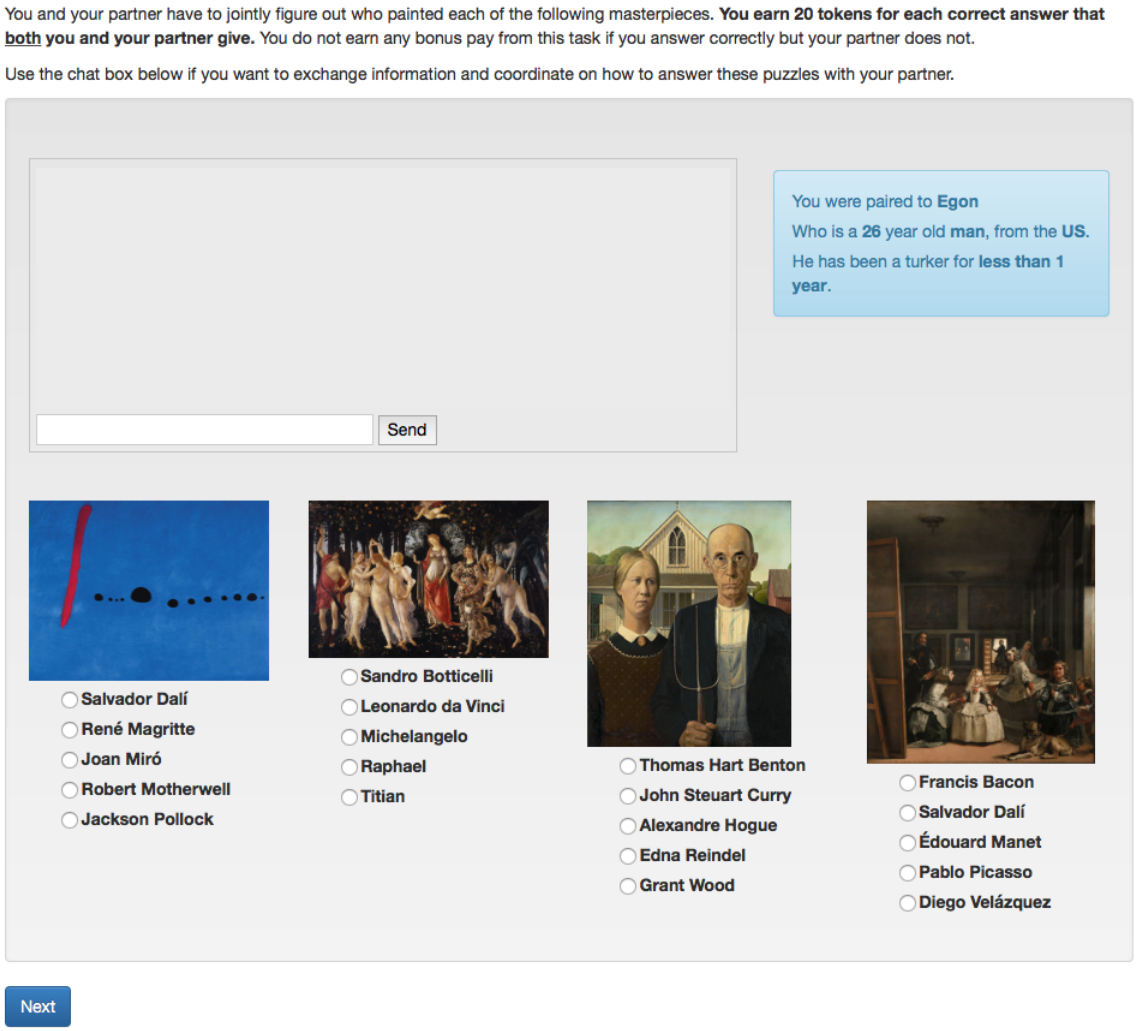
What is your gender? [drop-down list]

For how long have you been a turker? [drop-down list]

B.3.4 Page 3: Wait Page

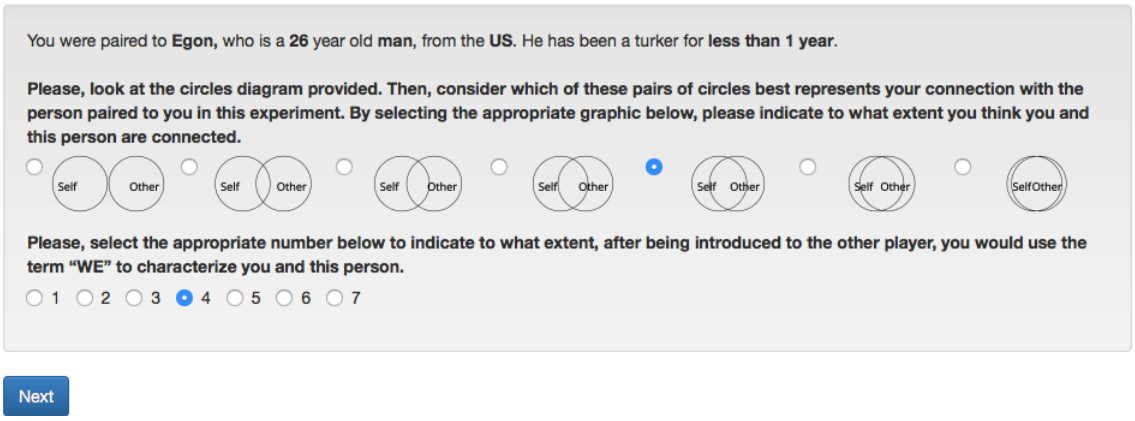
Please wait. Pairs are being formed.[3]

B.3.5 Page 4: Joint Problem Solving Task



[3] At this point of the experiment, each subject gets paired, randomly and anonymously, to another study participant.

B.3.6 Page 5: Oneness Elicitation



B.3.7 Page 6: Instructions for Donations

You will be able to engage in charitable giving by working on a simple assignment. Please carefully read the instructions below. Shortly, you will have the chance to familiarize yourself with this assignment in a training session. This will not affect your donation or payoffs. After the training, we will explain the payoffs for this task.

The assignment involves consecutively pressing the keys w e on your keyboard. You need to press the keys in this order. The keys are highlighted on the keyboard below. The software will display the number of successfully completed sequences.

You generate a donation to Doctors without Borders by completing a given number of sequences. A bar will indicate your progress towards this number.

In this example, you are asked to complete 100 keystroke sequences to generate a donation. Remember that this is just an example so that you can familiarize yourself with this assignment.

Please complete the training by pressing **w e** on your keyboard.

B.3.8 Page 7: Elicitation of Beliefs and Donations, and Treatment Assignment

You can choose to generate 50 tokens donations to Doctors Without Borders (DWB) by completing 100 keystroke sequences for each donation.

As incentive for yourself to complete donations, we offer a prize tied to the die face

you picked at the beginning of the experiment. For each donation you complete, you can earn 50 tokens. The player paired to you is offered the same incentive.[4]

[Name_other_player] is being [lucky/unlucky]. [He/She] picked number [n]. [His/Her] winning number is between [1 and 3/4 and 6]. [He/She] has [no chance/one chance in three] to win the 50 tokens prize for engaging in a donation, and has been informed of that.[5]

[Name_other_player] picked number [n]. [He/She] has one chance in six to win the 50 tokens prize for engaging in a donation, and is aware of that.[6]

You may be [lucky/unlucky]. You picked number [m] and your winning number is between [1 and 3/4 and 6]. You have [no chance/one chance in three] to win the 50 tokens prize for engaging in a donation.[7]

You picked number [m]. You have one chance in six to win the 50 tokens prize for engaging in a donation.[8]

You were paired to [Name_other_player], who is a [age_other_player] year old[man/woman] from the US. [He/She] has been a turker for [less than 1 year/1 year/2 years/more than 2 years].

[4] Text displayed only if incentives were available.
[5] Text displayed only if other player's incentives were either *Zero* or *High*.
[6] Text displayed only if other player's incentives were *Moderate*.
[7] Text displayed only if personal incentives were either *Zero* or *High*.
[8] Text displayed only if personal incentives were *Moderate*.

How many donations would you expect [Name_other_player] to complete? (you will earn 20 tokens if your guess is correct)

○ 0 Donations (0 tokens for DWB)

○ 1 Donation (50 tokens for DWB, and one chance in [six/three] to earn 50 tokens for [him/her]self)

○ 2 Donations (100 tokens for DWB, and one chance in [six/three] to earn 100 tokens for [him/her]self)

○ 3 Donations (150 tokens for DWB, and one chance in [six/three] to earn 150 tokens for [him/her]self)

○ 4 Donations (200 tokens for DWB, and one chance in [six/three] to earn 200 tokens for [him/her]self)

○ 5 Donations (250 tokens for DWB, and one chance in [six/three] to earn 250 tokens for [him/her]self)

○ 6 Donations (300 tokens for DWB, and one chance in [six/three] to earn 300 tokens for [him/her]self)

○ 7 Donations (350 tokens for DWB, and one chance in [six/three] to earn 350 tokens for [him/her]self)

○ 8 Donations (400 tokens for DWB, and one chance in [six/three] to earn 400 tokens for [him/her]self)

○ 9 Donations (450 tokens for DWB, and one chance in [six/three] to earn 450 tokens for [him/her]self)

○ 10 Donations (500 tokens for DWB, and one chance in [six/three] to earn 500 tokens for [him/her]self)

How many donations would you like to generate yourself?

○ 0 Donations (0 tokens for DWB)

○ 1 Donation (50 tokens for DWB, and one chance in [six/three] to earn 50 tokens for yourself)

○ 2 Donations (100 tokens for DWB, and one chance in [six/three] to earn 100 tokens for yourself)

○ 3 Donations (150 tokens for DWB, and one chance in [six/three] to earn 150 tokens for yourself)

○ 4 Donations (200 tokens for DWB, and one chance in [six/three] to earn 200 tokens for yourself)

○ 5 Donations (250 tokens for DWB, and one chance in [six/three] to earn 250 tokens for yourself)

○ 6 Donations (300 tokens for DWB, and one chance in [six/three] to earn 300 tokens for yourself)

○ 7 Donations (350 tokens for DWB, and one chance in [six/three] to earn 350 tokens for yourself)

○ 8 Donations (400 tokens for DWB, and one chance in [six/three] to earn 400 tokens for yourself)

○ 9 Donations (450 tokens for DWB, and one chance in [six/three] to earn 450 tokens for yourself)

○ 10 Donations (500 tokens for DWB, and one chance in [six/three] to earn 500 tokens for yourself[a])

[a] Text displayed only if private incentives are available with positive ex-interim probability.

B.3.9 Page 8: Donation Task

You have chosen to make [D] donations. For this you will have to complete [D×100] keystroke sequences to generate these donations

Please complete the donation to Doctors without Borders by pressing w e on your keyboard.

B.3.10 Page 9: Short Questionnaire

Thank you for completing the donation task. Please fill out the short questionnaire below and then go to the next page to review payoffs and complete the HIT.

You and your partner could earn 20 tokens for guessing correctly how many donations the other did. Aside from the guessing question, was it clear to you that the number of donations that YOU made was not directly affecting the payoff of the other player? [Yes/No]

You and your partner could earn 20 tokens for guessing correctly how many donations the other did. Aside from the guessing question, was it clear to you that the number of donations that the OTHER made was not directly affecting your payoff? [Yes/No]

Did you realize that the amount donated to charity was increasing in the donations that both you and the other player made? [Yes/No]

In choosing how many donations to make, were you influenced by the thought of the number of donations the other person was going to make? [Yes/No]

Expecting that the other person could make more donations, makes you want to donate [More/Less/Indifferent]

In other contexts, when you are about to make a charitable donation, do you ever consider whether and how much other people are contributing to the same cause? [Always/Very often/Sometimes/Rarely/Never]

In other contexts, when you are about to make a charitable donation, expecting that other people could make more donations, makes you want to donate [More/Less/Indifferent]

Please recall the screen where you chose how many donations to make. What

were the chances YOU had to win the lottery for participating in the donation task? [No chances/One chance in six/One chance in three/Cannot recall]

Please recall the screen where you chose how many donations to make. What were the chances the OTHER player had to win the lottery for participating in the donation task? [No chances/One chance in six/One chance in three/Cannot recall][9]

[9] Questions displayed only if incentives were available.

Appendix C

Appendix to Chapter 3

C.1 Appendix: Proofs

Proof of Proposition 1. The proposition is composed of two statements.

First statement: *"A dual market for donations increases contributions compared to a single market where no incentives are available."*

When actions are private, the utility of any agent i can be re-written as

$$U_i(d,y) = \begin{cases} [a_i(B-y)+y-c]d, & \text{Dual Market: } y \in \{0,\tilde{y}\} \\ [a_iB-c]d, & \text{Single Market - No Incentives } y=0 \end{cases}$$

Availability of incentives $\tilde{y} > 0$ does not affect donation behavior of highly altruistic agents ($a_i > 1$), who can choose to turn down the incentive, gaining utility

$$a_iB - c > a_i(B-\tilde{y}) + \tilde{y} - c.$$

At the same time, the availability of incentives get agents for whom

$$a_iB - c < 0 < a_i(B-\tilde{y}) + \tilde{y} - c$$

involved in the donation.

When actions take place in public, the same as above applies for image-indifferent agents. Image-concerned agents will now focus instead on taking the action that sends the best possible signal about their degree of altruism. Independence in the distribution of the degree of altruism and image concern implies that image-concerned agents would never refrain from donating, as doing so would send the worst possible signal about their degree of altruism.

Second statement: *"Compared to a single market where conditional incentives are au-*

tomatic and cannot be turned down, allowing to turn down incentives reduces the cost of collection without affecting the number of donations."

When actions are private, the utility of any agent i can be re-written as

$$U_i(d,y) = \begin{cases} [a_i(B-y)+y-c]d, & \text{Dual Market: } y \in \{0,\tilde{y}\} \\ [a_i(B-\tilde{y})+\tilde{y}-c]d, & \text{Single Market - with Incentives} \end{cases}$$

Define the share of highly altruistic agents as $s(a) = Pr(a_i > 1)$. Because $B > c$, a $s(a)$ share of agents would donate irrespective of the availability of incentives, even though their intrinsic motivation to donate is partially crowded out in a single market with incentives. Allowing agents, in a dual market, to sort out of incentives un-does the described crowding out of intrinsic motivation to donate and reduces the average cost of collection.

When actions take place in public, the same as in private applies for image-indifferent agents. For image concerned agents, we need to show that participation is unaffected by the possibility of turning down incentives. Therefore, we need to show that in neither a single incentivized market nor in a dual market image concerned agents want to abstein from donating. The proof goes by contradiction.

In a dual market, suppose there exists a pure strategy equilibrium in which all image concerned agents were to not donate. Any one of these agents could deviate from the equilibrium by donating and turning down the compensation to mimic the most altruistic image indifferent agents. Such deviation would improve the reputation of this agent, hence her utility. A contradiction.

Similarly, in the single incentivized market the profitable deviation is represented by the reputational gain of donating with incentives.

□

C.2 Appendix: Additional Tables

Table C.2.1: Poisson Regression for Total Individual Donations: Semi-Elasticities (Coefficient Estimates and Standard Errors in Parentheses)

Dependent variable:	# of donations over the three rounds				
	(1)	(2)	(3)	(4)	(5)
	a) Treatments				
Paid&Choose *(baseline Not Paid)*	0.232*	0.232*	0.261**	0.260**	0.294**
	(0.133)	(0.132)	(0.121)	(0.121)	(0.120)
Public *(baseline Private)*	0.290**	0.298**	0.299**	0.301**	0.320**
	(0.141)	(0.140)	(0.132)	(0.131)	(0.129)
Paid&Choose × Public	-0.004	-0.008	0.021	0.020	-0.010
	(0.163)	(0.162)	(0.152)	(0.152)	(0.150)
	b) Controls				
Female		0.151**		0.048	0.019
		(0.075)		(0.070)	(0.070)
DG: Dictator kept			-0.062***	-0.061***	-0.057***
			(0.009)	(0.009)	(0.009)
Other controls	No	No	No	No	Yes
Observations	329	329	329	329	329

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Notes: Standard errors are clustered at individual level. *NOT PAID* is the base market design treatment. *PRIVATE* is the base visibility treatment. Other controls include age, chosen charity, and individual rating of chosen charity.

Table C.2.2: Random Effects Regressions: Relative Risk Ratios (Coefficient Estimates and Standard Errors in Parentheses)

Dependent variable:	Incentive Choice				
	(1)	(2)	(3)	(4)	(5)
	a) Treatment				
Public	1.747	1.652	1.862	1.705	2.229
	(1.278)	(1.155)	(1.439)	(1.257)	(1.585)
	b) Controls				
Female		0.719		0.552	0.601
		(0.531)		(0.420)	(0.456)
DG: Tokens kept			0.928	0.915	0.934
			(0.104)	(0.102)	(0.104)
Other controls	No	No	No	No	Yes
Observations	378	378	378	378	378

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$ for relative risk ratios different from unity.

Notes: Standard errors are clustered at the individual level. *PRIVATE* is the base visibility treatment. The incentive choice dependent variable only applies to the 126 subjects in *CHOOSE* treatment over three rounds. Incentive choice takes value ”0” if subject skips, ”1” if participates unpaid, and ”2” if participates paid to the donation task in a given round. The table reports relative risk ratio for outcome ”1” unpaid participation and base outcome ”2” paid participation.

Table C.2.3: Poisson Regression for Total Individual Donations: Semi-Elasticities (Coefficient Estimates and Standard Errors in Parentheses)

	Incentive Treatment Subsamples		
	Not paid	Paid	Choose
	(1)	(2)	(3)
a) Gender dummy × visibility treatment			
Public	0.483*	0.342	0.713***
	(0.253)	(0.210)	(0.268)
Female	0.230	0.086	0.357
	(0.242)	(0.196)	(0.258)
Public × Female	-0.285	0.026	-0.584*
	(0.293)	(0.251)	(0.315)
b) Controls			
DG: Tokens kept	-0.050***	-0.053***	-0.083***
	(0.015)	(0.015)	(0.019)
Observations	93	110	126

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: Standard errors are clustered at individual level. *NOT PAID* is the base market design treatment. *PRIVATE* is the base visibility treatment. DG refers to the dictator game, in which we gave 20 experimental tokens to subjects and asked them how many they would like to keep.

Table C.2.4: Poisson Regression for Total Individual Donations (Coefficient Estimates and Standard Errors in Parentheses)

Dependent variable:	# of donations over the three rounds					
	Semi-elasticities			Average marginal effects		
	(1)	(2)	(3)	(4)	(5)	(6)
	a) Treatments					
Paid	0.149	0.183	0.205	0.268**	0.322***	0.333***
	(0.157)	(0.142)	(0.140)	(0.129)	(0.121)	(0.118)
Choose	0.294**	0.318**	0.363***	0.409***	0.476***	0.512***
	(0.145)	(0.131)	(0.131)	(0.128)	(0.117)	(0.118)
Public	0.290**	0.301**	0.320**	0.451***	0.496***	0.492***
	(0.141)	(0.131)	(0.129)	(0.107)	(0.098)	(0.097)
Paid × Public	0.056	0.065	0.043			
	(0.189)	(0.175)	(0.172)			
Choose × Public	-0.044	-0.008	-0.043			
	(0.179)	(0.165)	(0.163)			
	b) Controls					
Female		0.040	0.010		0.064	0.016
		(0.069)	(0.070)		(0.109)	(0.111)
DG: Tokens kept		-0.062***	-0.057***		-0.097***	-0.091***
		(0.009)	(0.009)		(0.013)	(0.013)
Other controls	No	No	Yes	No	No	Yes
Observations	329	329	329	329	329	329

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Notes: Standard errors are clustered at individual level. *NOT PAID* is the base market design treatment. *PRIVATE* is the base visibility treatment. DG refers to the dictator game, in which we gave 20 experimental tokens to subjects and asked them how many they would like to keep. Other controls include age, chosen charity, and individual rating of chosen charity.

References

Mercier, Hugo and Dan Sperber (2011). "Why do humans reason? Arguments for an argumentative theory". In: *Behavioral and brain sciences* 34.2, pp. 57–74.

Von Hippel, William and Robert Trivers (2011). "The evolution and psychology of self-deception". In: *Behavioral and Brain Sciences* 34.1, pp. 1–16.

Kurzban, Robert (2012). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton University Press.

Mercier, Hugo (2016). "The argumentative theory: Predictions and empirical evidence". In: *Trends in Cognitive Sciences* 20.9, pp. 689–700.

Bénabou, Roland, Armin Falk, and Jean Tirole (2019). "Narratives, Imperatives and Moral Reasoning". In:

Milgrom, Paul R (1981). "Good news and bad news: Representation theorems and applications". In: *The Bell Journal of Economics*, pp. 380–391.

Crawford, Vincent P and Joel Sobel (1982). "Strategic information transmission". In: *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian persuasion". In: *American Economic Review* 101.6, pp. 2590–2615.

DellaVigna, Stefano and Matthew Gentzkow (2010). "Persuasion: Empirical Evidence". In: *Annual Review of Economics* 2.1, pp. 643–669.

Tappin, Ben M, Gordon Pennycook, and David Rand (2019). "Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often prevent causal inference". In:

Festinger, Leon and James M Carlsmith (1959). "Cognitive consequences of forced compliance." In: *The journal of abnormal and social psychology* 58.2, p. 203.

Elliot, Andrew J and Patricia G Devine (1994). "On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort." In: *Journal of personality and social psychology* 67.3, p. 382.

Chen, Zhuoqiong Charlie and Tobias Gesche (2017). "Persistent bias in advice-giving". In: *University of Zurich, Department of Economics, Working Paper* 228.

Gneezy, Uri et al. (2020). "Bribing the self". In: *Games and Economic Behavior* 120, pp. 311–324.

Smith, Megan K., Robert Trivers, and William von Hippel (Dec. 2017). "Self-deception facilitates interpersonal persuasion". In: *Journal of Economic Psychology* 63, pp. 93–101.

Schwardmann, Peter and Joel van der Weele (2019). *Deception and Self-Deception*. Rationality and Competition Discussion Paper Series.

Solda, Alice et al. (2019). *Strategically delusional*. Tech. rep. QUT Business School.

Janis, Irving L. and Bert T. King (1954). "The influence of role playing on opinion change". In: *Journal of Abnormal and Social Psychology* 49.2, pp. 211–218.

O'Neill, Patrick and Diane E. Levings (1979). "Inducing biased scanning in a group setting to change attitudes toward bilingualism and capital punishment." In: *Journal of Personality and Social Psychology* 37.8, pp. 1432–1438.

Engel, Christoph and Andreas Glöckner (2013). "Role-Induced Bias in Court: An Experimental Analysis". In: *Journal of Behavioral Decision Making* 26.3, pp. 272–284.

Babcock, Linda et al. (1995). "Biased judgments of fairness in bargaining". In: *The American Economic Review* 85.5, pp. 1337–1343.

List, JA (2003). "Does market experience eliminate market anomalies?" In: *The Quarterly Journal of Economics* 118.1, pp. 41–71.

Kunda, Ziva (1990). "The case for motivated reasoning." In: *Psychological bulletin* 108.3, p. 480.

Bénabou, Roland and Jean Tirole (2016). "Mindful economics: The production, consumption, and value of beliefs". In: *Journal of Economic Perspectives* 30.3, pp. 141–64.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber (2016). "Motivated Bayesians: Feeling moral while acting egoistically". In: *Journal of Economic Perspectives* 30.3, pp. 189–212.

Di Tella, Rafael, S. Galiant, and E. Schargrodsky (2007). "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters". In: *The Quarterly Journal of Economics* 122.1, pp. 209–241.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey (2013). "Optimal expectations and limited medical testing: Evidence from huntington disease". In: *American Economic Review* 103.2, pp. 804–830.

Schlag, Karl H, James Tremewan, and Joël J Van der Weele (2015). "A penny for your thoughts: A survey of methods for eliciting beliefs". In: *Experimental Economics* 18.3, pp. 457–490.

Lord, Charles G, Lee Ross, and Mark R Lepper (1979). "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." In: *Journal of personality and social psychology* 37.11, p. 2098.

Sunstein, Cass R. (2002). "The Law of Group Polarization". In: *Journal of Political Philosophy* 10.2, pp. 175–195.

Taber, Charles S. and Milton Lodge (2006). "Motivated Skepticism in the Evaluation of Political Beliefs". In: *American Journal of Political Science* 50.3, pp. 755–769.

Kahan, Dan (Dec. 2015). *The Politically Motivated Reasoning Paradigm*. SSRN Scholarly Paper ID 2703011. Rochester, NY: Social Science Research Network.

Fryer, Roland G., Philipp Harms, and Matthew O. Jackson (2018). "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization". In: *Journal of the European Economic Association*.

Gennaioli, Nicola and Guido Tabellini (2019). "Identity, Beliefs, and Political Conflict". In:

Eliaz, Kfir and Rani Spiegler (2018). "A Model of Competing Narratives". In:

Foerster, Manuel and Joel J van der Weele (2018). "Persuasion, justification and the communication of social impact". In:

Bullock, John G et al. (2015). "Partisan Bias in Factual Beliefs about Politics". In: *Quarterly Journal of Political Science* 10.4, pp. 519–578.

Habermas, Jürgen (1984). *The theory of communicative action*. Vol. 1. Beacon Press.

Elster, Jon (1998). *Deliberative Democracy*. Cambridge Studies in the Theory of Democracy. Cambridge University Press.

Gutmann, Amy and Dennis Thompson (2004). *Why Deliberative Democracy?* Student edition. Princeton University Press. ISBN: 9780691120195.

Kuhn, Deanna, Victoria Shaw, and Mark Felton (1997). "Effects of Dyadic Interaction on Argumentive Reasoning". In: *Cognition and Instruction* 15.3, pp. 287–315.

Thompson, Dennis F. (2008). "Deliberative democratic theory and empirical political science". In: *Annu. Rev. Polit. Sci.* 11, pp. 497–520.

Mercier, Hugo and Hélène Landemore (2012). "Reasoning is for arguing: Understanding the successes and failures of deliberation". In: *Political Psychology* 33.2, pp. 243–258.

Hossain, Tanjim and Ryo Okui (2013). "The binarized scoring rule". In: *Review of Economic Studies* 80.3, pp. 984–1001.

Schlag, Karl H. and Joël J. Van der Weele (2013). "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality". In: *Theoretical Economics Letters* 3.1, pp. 38–42.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg (2017). "Culture, ethnicity, and diversity". In: *American Economic Review* 107.9, pp. 2479–2513.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray (Nov. 2004). "Polarization: Concepts, Measurement, Estimation". In: *Econometrica* 72.6, pp. 1737–1772.

Mutz, Diana C. (Nov. 2007). "Effects of "In-Your-Face" Television Discourse on Perceptions of a Legitimate Opposition". In: *American Political Science Review* 101.4, pp. 621–635.

Druckman, James N. and Arthur Lupia (2016). "Preference Change in Competitive Political Environments". In: *Annual Review of Political Science* 19.1, pp. 13–31. eprint: https://doi.org/10.1146/annurev-polisci-020614-095051.

Juslin, Peter, Anders Winman, and Patrik Hansson (2007). "The naive intuitive statistician: a naive sampling model of intuitive confidence intervals." In: *Psychological review* 114.3, p. 678.

Barron, Kai, Steffen Huck, and Philippe Jehiel (2019). *Everyday econometricians: Selection neglect and overoptimism when learning from others*. Tech. rep. WZB Discussion Paper.

Tversky, Amos and Daniel Kahneman (1973). "Availability: A heuristic for judging frequency and probability". In: *Cognitive Psychology* 5.2, pp. 207–232.

Vinokur, Amiram and Eugene Burstein (1974). "Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach". In: *Journal of Personality and Social Psychology* 29.3, pp. 305–315.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). "Identification, inference and sensitivity analysis for causal mediation effects". In: *Statistical science*, pp. 51–71.

Exley, Christine L and Judd B Kessler (2019). *Motivated errors*. Tech. rep. National Bureau of Economic Research.

Cessie, Saskia le et al. (2012). "Quantification of bias in direct effects estimates due to different types of measurement error in the mediator". In: *Epidemiology*, pp. 551–560.

Imai, Kosuke, Luke Keele, and Dustin Tingley (2010). "A general approach to causal mediation analysis." In: *Psychological methods* 15.4, p. 309.

Petersen, Michael Bang et al. (2013). "Motivated reasoning and political parties: Evidence for increased processing in the face of party cues". In: *Political Behavior* 35.4, pp. 831–854.

Falk, Armin and Florian Zimmermann (2016). "Consistency as a Signal of Skills". In: *Management Science* 63.7, pp. 2197–2210.

Bénabou, Roland and Jean Tirole (2002). "Self-Confidence and Personal Motivation". In: *The Quarterly Journal of Economics* 117.3, pp. 871–915.

Felton, Mark, Amanda Crowell, and Tina Liu (2015). "Arguing to agree: Mitigating my-side bias through consensus-seeking dialogue". In: *Written Communication* 32.3, pp. 317–331.

Perkins, David (2019). "Learning to reason: The influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues". In: *Judgment and Decision Making* 14.6, p. 624.

Fiorina, Morris P and Samuel J Abrams (2008). "Political polarization in the American public". In: *Annu. Rev. Polit. Sci.* 11, pp. 563–588.

Bail, Christopher A. et al. (2018). "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37, pp. 9216–9221. eprint: https://www.pnas.org/content/115/37/9216.full.pdf.

Mullainathan, Sendhil and Ebonya Washington (2009). "Sticking with your vote: Cognitive dissonance and political attitudes". In: *American Economic Journal: Applied Economics* 1.1, pp. 86–111.

Gould, Eric D. and Esteban F. Klor (2019). "Party hacks and true believers: The effect of party affiliation on political preferences". In: *Journal of Comparative Economics*.

Gal, David and Derek D Rucker (2010). "When in doubt, shout! Paradoxical influences of doubt on proselytizing". In: *Psychological Science* 21.11, pp. 1701–1707.

Linnainmaa, Juhani T, Brian Melzer, and Alessandro Previtero (2018). "The misguided beliefs of financial advisors". In: *Kelley School of Business Research Paper* 18-9.

Cheng, Ing-Haw, Sahil Raina, and Wei Xiong (2015). "Wall Street and the Housing Bubble: Bad Incentives, Bad Models, or Bad Luck?" In: *American Economic Review* 104.9, pp. 2797–2829.

Trivers, Robert (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books (AZ).

Nickerson, Raymond S. (1998). "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of General Psychology* 2.2, pp. 175–220.

Benjamin, Daniel J (2019). "Errors in probabilistic reasoning and judgment biases". In: *Handbook of Behavioral Economics-Foundations and Applications 2*, p. 69.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007). "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness". In: *Economic Theory* 33.1, pp. 67–80.

Exley, Christine L (2015). "Excusing selfishness in charitable giving: The role of risk". In: *The Review of Economic Studies* 83.2, pp. 587–628.

Di Tella, Rafael et al. (2015). "Conveniently Upset: Avoiding Altruism by Distorting Beliefs About Others". In: *American Economic Review* 105.11, pp. 3416–3442.

Simler, Kevin and Robin Hanson (2017). *The elephant in the brain: Hidden motives in everyday life*. Oxford University Press.

Frey, Bruno S. and Stephan Meier (2004). "Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment". In: *The American Economic Review* 94.5, pp. 1717–1722.

Bursztyn, Leonardo et al. (2014). "Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions". In: *Econometrica* 82.4, pp. 1273–1301.

Bapna, Ravi and Akhmed Umyarov (Apr. 2015). "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks". In: *Management Science* 61.8, pp. 1902–1920.

Cantoni, Davide et al. (Jan. 2017). *Are Protests Games of Strategic Complements or Substitutes? Experimental Evidence from Hong Kong's Democracy Movement*. Working Paper 23110. DOI: 10.3386/w23110. National Bureau of Economic Research.

Drago, Francesco, Friederike Mengel, and Christian Traxler (2020). "Compliance behavior in networks: Evidence from a field experiment". In: *American Economic Journal: Applied Economics* 12.2, pp. 96–133.

Aral, Sinan and Christos Nicolaides (Apr. 2017). "Exercise contagion in a global social network". In: *Nature Communications* 8, ncomms14753.

Bernheim, B. Douglas (1994). "A Theory of Conformity". In: *Journal of Political Economy* 102.5, pp. 841–877.

Akerlof, George A. (1997). "Social Distance and Social Decisions". In: *Econometrica* 65.5, pp. 1005–1027.

Becker, Gary S. (Nov. 1974). "A Theory of Social Interactions". In: *Journal of Political Economy* 82.6, pp. 1063–1093.

Andreoni, James (Dec. 1989). "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence". In: *Journal of Political Economy* 97.6, pp. 1447–1458.

Andreoni, James (June 1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". In: *The Economic Journal* 100.401, p. 464.

Chen, Yan and Sherry Xin Li (2009). "Group identity and social preferences". In: *American Economic Review* 99.1, pp. 431–57.

Chen, Roy and Yan Chen (Oct. 2011). "The Potential of Social Identity for Equilibrium Selection". In: *American Economic Review* 101.6, pp. 2562–2589.

Cialdini, Robert B. et al. (1997). "Reinterpreting the empathy–altruism relationship: When one into one equals oneness." In: *Journal of Personality and Social Psychology* 73.3, pp. 481–494.

Fuster, Andreas and Stephan Meier (Oct. 2009). "Another Hidden Cost of Incentives: The Detrimental Effect on Norm Enforcement". In: *Management Science* 56.1, pp. 57–70.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani (2017). "The Morale Effects of Pay Inequality". In: *The Quarterly Journal of Economics*.

Cason, Timothy N. and Vai-Lam Mui (June 1998). "Social Influence in the Sequential Dictator Game". In: *Journal of Mathematical Psychology* 42.2, pp. 248–265.

Bohnet, Iris and Richard Zeckhauser (Oct. 2004). "Social Comparisons in Ultimatum Bargaining". In: *Scandinavian Journal of Economics* 106.3, pp. 495–510.

Eckel, Catherine C. and Rick K. Wilson (Sept. 2007). "Social learning in coordination games: does status matter?" In: *Experimental Economics* 10.3, pp. 317–329.

Krupka, Erin and Roberto A. Weber (June 2009). "The focusing and informational effects of norms on pro-social behavior". In: *Journal of Economic Psychology* 30.3, pp. 307–320.

Servátka, Maroš (2009). "Separating reputation, social influence, and identification effects in a dictator game". In: *European Economic Review* 53.2, pp. 197–209.

Duffy, John and Tatiana Kornienko (May 2010). "Does competition affect giving?" In: *Journal of Economic Behavior & Organization* 74.1, pp. 82–103.

Bigenho, Jason and Seung-Keun Martinez (2019). *Social Comparisons in Peer Effects*. Tech. rep. Technical report, UCSD.

Shang, Jen and Rachel Croson (Oct. 2009). "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods". In: *The Economic Journal* 119.540, pp. 1422–1439.

Chen, Yan et al. (Sept. 2010). "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens". In: *American Economic Review* 100.4, pp. 1358–1398.

Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler (June 2013). "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information". In: *Journal of the European Economic Association* 11.3, pp. 634–660.

Bruhin, Adrian et al. (2020). "Spillovers of Prosocial Motivation: Evidence from an Intervention Study on Blood Donors". In: *Journal of Health Economics* 70, p. 102244.

Lahno, Amrei M and Marta Serra-Garcia (2015). "Peer effects in risk taking: Envy or conformity?" In: *Journal of Risk and Uncertainty* 50.1, pp. 73–95.

Gilchrist, Duncan Sheppard and Emily Glassberg Sands (2016). "Something to talk about: Social spillovers in movie consumption". In: *Journal of Political Economy* 124.5, pp. 1339–1382.

Eckel, Catherine C and Philip J Grossman (Mar. 2003). "Rebate versus matching: does how we subsidize charitable contributions matter?" In: *Journal of Public Economics* 87.3–4, pp. 681–701.

Landry, Craig E et al. (2006). "Toward an understanding of the economics of charity: Evidence from a field experiment". In: *The Quarterly journal of economics* 121.2, pp. 747–782.

Huck, Steffen, Imran Rasul, and Andrew Shephard (May 2015). "Comparing Charitable Fundraising Schemes: Evidence from a Natural Field Experiment and a Structural Model". In: *American Economic Journal: Economic Policy* 7.2, pp. 326–369.

Meer, Jonathan (2017). "Does fundraising create new giving?" In: *Journal of Public Economics* 145, pp. 82–93.

Perez-Truglia, Ricardo and Guillermo Cruces (2017). "Partisan interactions: Evidence from a field experiment in the united states". In: *Journal of Political Economy* 125.4, pp. 1208–1243.

Gneezy, Uri and Aldo Rustichini (2000a). "A Fine Is a Price". In: *Journal of Legal Studies* 29, pp. 1–18.

— (2000b). "Pay Enough or Don't Pay at All". In: *The Quarterly Journal of Economics* 115.3, pp. 791–810.

Topa, Giorgio (Apr. 2001). "Social Interactions, Local Spillovers and Unemployment". In: *The Review of Economic Studies* 68.2, pp. 261–295.

Leider, Stephen et al. (Nov. 2009). "Directed Altruism and Enforced Reciprocity in Social Networks". In: *The Quarterly Journal of Economics* 124.4, pp. 1815–1851.

Bond, Robert M. et al. (Sept. 2012). "A 61-million-person experiment in social influence and political mobilization". In: *Nature* 489.7415, pp. 295–298.

Dimant, Eugen (2018). *Contagion of Pro-and Anti-Social Behavior Among Peers and the Role of Social Proximity*. Tech. rep.

Kessler, Judd B. (Dec. 2017). "Announcements of Support and Public Good Provision". In: *American Economic Review* 107.12.

Gächter, Simon, Chris Starmer, and Fabio Tufano (June 2015). "Measuring the Closeness of Relationships: A Comprehensive Evaluation of the 'Inclusion of the Other in the Self' Scale". In: *PLOS ONE* 10.6, e0129478.

Aron, Arthur, Elaine N Aron, and Danny Smollan (1992). "Inclusion of other in the self scale and the structure of interpersonal closeness." In: *Journal of personality and social psychology* 63.4, p. 596.

Ariely, Dan, Anat Bracha, and Stephan Meier (Feb. 2009). "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially". In: *American Economic Review* 99.1, pp. 544–555.

Meyer, Christian Johannes and Egon Tripodi (Oct. 2017). *Sorting into Incentives for Prosocial Behavior*. SSRN Scholarly Paper ID 3058195. Rochester, NY: Social Science Research Network.

DellaVigna, Stefano and Devin Pope (Aug. 2016). *Predicting Experimental Results: Who Knows What?* Working Paper 22566. DOI: 10.3386/w22566. National Bureau of Economic Research.

— (2017). "What motivates effort? Evidence and expert forecasts". In: *The Review of Economic Studies* 85.2, pp. 1029–1069.

Ellingsen, Tore and Magnus Johannesson (2008). "Pride and prejudice: The human side of incentive theory". In: *American economic review* 98.3, pp. 990–1008.

DellaVigna, Stefano (2018). *Structural behavioral economics*. Tech. rep. National Bureau of Economic Research.

Sliwka, Dirk (2007). "Trust as a signal of a social norm and the hidden costs of incentive schemes". In: *American Economic Review* 97.3, pp. 999–1012.

Kelman, HC (1961). "Processes of opinion change". In: *Public Opinion Quarterly* 25.

Gioia, Francesca (2017). "Peer effects on risk behaviour: the importance of group identity". In: *Experimental Economics* 20.1, pp. 100–129.

Benabou, Roland and Jean Tirole (Dec. 2006). "Incentives and Prosocial Behavior". In: *American Economic Review* 96.5, pp. 1652–1678.

Dutta, Rohan, David K Levine, and Salvatore Modica (2018). *Peer Monitoring, Ostracism and the Internalization of Social Norms*. Tech. rep. David K. Levine.

Chen, Daniel L., Martin Schonger, and Chris Wickens (Mar. 2016b). "oTree-An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.

Hauser, David J. and Norbert Schwarz (Mar. 2016). "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants". In: *Behavior Research Methods* 48.1, pp. 400–407.

Hara, Kotaro et al. (2018). "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, p. 449.

Goette, Lorenz, David Huffman, and Stephan Meier (Feb. 2012). "The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups". In: *American Economic Journal: Microeconomics* 4.1, pp. 101–115.

Chen, Yan et al. (Mar. 2014). "Which hat to wear? Impact of natural identities on coordination and cooperation". In: *Games and Economic Behavior* 84.Supplement C, pp. 58–86.

Marmaros, David and Bruce Sacerdote (2006). "How do friendships form?" In: *The Quarterly Journal of Economics* 121.1, pp. 79–119.

Goette, Lorenz, David Huffman, and Stephan Meier (2006). "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups". In: *The American Economic Review* 96.2, pp. 212–216.

Burks, Stephen V. et al. (May 2009). "Cognitive skills affect economic preferences, strategic behavior, and job attachment". In: *Proceedings of the National Academy of Sciences* 106.19, pp. 7745–7750.

Jones, Daniel and Sera Linardi (Mar. 2014). "Wallflowers: Experimental Evidence of an Aversion to Standing Out". In: *Management Science* 60.7, pp. 1757–1771.

Mas, Alexandre and Enrico Moretti (2009). "Peers at work". In: *American Economic Review* 99.1, pp. 112–45.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010). "Social incentives in the workplace". In: *The review of economic studies* 77.2, pp. 417–458.

Smith, Alexander (Sept. 2013). "Estimating the causal effect of beliefs on contributions in repeated public good games". In: *Experimental Economics* 16.3, pp. 414–425.

Costa-Gomes, Miguel A., Steffen Huck, and Georg Weizsäcker (Nov. 2014). "Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect". In: *Games and Economic Behavior* 88.Supplement C, pp. 298–309.

Foltz, Jeremy D and Kweku A Opoku-Agyemang (2015). "Do higher salaries lower petty corruption?" In: *A Policy Experiment on West Africa?s Highways*.

Ferraro, Paul J and Michael K Price (2013). "Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment". In: *Review of Economics and Statistics* 95.1, pp. 64–73.

Allcott, Hunt and Todd Rogers (2014). "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation". In: *American Economic Review* 104.10, pp. 3003–37.

Frey, Bruno S. and Felix Oberholzer-Gee (1997). "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out". In: *The American Economic Review* 87.4, pp. 746–755.

Bowles, Samuel and Sandra Polanía-Reyes (June 2012). "Economic Incentives and Social Preferences: Substitutes or Complements?" In: *Journal of Economic Literature* 50.2, pp. 368–425.

Deci, Edward L. (1975). *Intrinsic Motivation*. Boston, MA: Springer US.

— (1971). "Effects of externally mediated rewards on intrinsic motivation". In: *Journal of Personality and Social Psychology* 18.1, pp. 105–115.

Titmuss, Richard M. (1971). *The Gift Relationship: From Human Blood to Social Policy*. New York, NY: Pantheon Books.

Exley, Christine (Mar. 2017). "Incentives for Prosocial Behavior: The Role of Reputations". In: *Management Science*.

Frey, Bruno S. and Lorenz Goette (1999). *Does Pay Motivate Volunteers?* IEW - Working Paper 007. Institute for Empirical Research in Economics - University of Zurich.

Lacetera, Nicola, Mario Macis, and Sarah S. Stith (Jan. 2014). "Removing financial barriers to organ and bone marrow donation: The effect of leave and tax legislation in the U.S." In: *Journal of Health Economics* 33, pp. 43–56.

Lacetera, Nicola, Mario Macis, and Robert Slonim (Feb. 2012). "Will There Be Blood? Incentives and Displacement Effects in Pro-Social Behavior". In: *American Economic Journal: Economic Policy* 4.1, pp. 186–223.

— (May 2014). "Rewarding Volunteers: A Field Experiment". In: *Management Science* 60.5, pp. 1107–1129.

Boulware, L. E. et al. (Nov. 2006). "Public Attitudes Toward Incentives for Organ Donation: A National Study of Different Racial/Ethnic and Income Groups". In: *American Journal of Transplantation* 6.11, pp. 2774–2785.

Becker, Gary S. and Julio Jorge Elias (2007). "Introducing Incentives in the Market for Live and Cadaveric Organ Donations". In: *The Journal of Economic Perspectives* 21.3, pp. 3–24.

Lacetera, Nicola (Sept. 2016). *Incentives and Ethics in the Economics of Body Parts*. Working Paper 22673. Cambridge, MA: National Bureau of Economic Research.

Niessen-Ruenzi, Alexandra, Martin Weber, and David M. Becker (2015). *To pay or not to pay – Evidence from whole blood donations in Germany*. mimeo.

The Lancet (June 2005). "Blood Supply and Demand". In: *The Lancet* 365.9478, p. 2151.

World Health Organization (2009). *The Melbourne Declaration on 100% Voluntary Non-Remunerated Donation of Blood and Blood Components*.

Council of Europe (1995). *Recommendation No. R (95) 14 of the Committee of Ministers to Member States on the Protection of the Health of Donors and Recipients in the Area of Blood Transfusion. Adopted by the Committee of Ministers on 12 October 1995 at the 545th Meeting of the Ministers' Deputies.* Tech. rep. Strasbourg: Council of Europe.

Carpenter, Jeffrey and Caitlin Knowles Myers (Dec. 2010). "Why volunteer? Evidence on the role of altruism, image, and incentives". In: *Journal of Public Economics* 94.11–12, pp. 911–920.

Mellstrom, Carl and Magnus Johannesson (June 2008). "Crowding Out in Blood Donation: Was Titmuss Right?" In: *Journal of the European Economic Association* 6.4, pp. 845–863.

Lacetera, Nicola, Mario Macis, and Robert Slonim (May 2013). "Economic Rewards to Motivate Blood Donations". In: *Science* 340.6135, pp. 927–928.

Chao, Matthew (June 2017). "Demotivating incentives and motivation crowding out in charitable giving". In: *Proceedings of the National Academy of Sciences*, pp. 7301–7306.

Filiz-Ozbay, Emel and Erkut Y. Ozbay (June 2014). "Effect of an audience in public goods provision". In: *Experimental Economics* 17.2, pp. 200–214.

Lacetera, Nicola and Mario Macis (Nov. 2010). "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme". In: *Journal of Economic Behavior & Organization* 76.2, pp. 225–237.

Bursztyn, Leonardo and Robert Jensen (Aug. 2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". In: *Annual Review of Economics* 9, pp. 131–153.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles (Nov. 2001). "Costly Signaling and Cooperation". In: *Journal of Theoretical Biology* 213.1, pp. 103–119.

Smith, Eric Alden and Rebecca L. Bliege Bird (July 2000). "Turtle hunting and tombstone opening: public generosity as costly signaling". In: *Evolution and Human Behavior* 21.4, pp. 245–261.

Eagly, Alice H. and Maureen Crowley (1986). "Gender and helping behavior: A meta-analytic review of the social psychological literature". In: *Psychological Bulletin* 100.3, pp. 283–308.

Iredale, Wendy, Mark Van Vugt, and Robin Dunbar (July 2008). "Showing Off in Humans: Male Generosity as a Mating Signal". In: *Evolutionary Psychology* 6.3, pp. 386–392.

Barclay, Pat (Feb. 2010). "Altruism as a courtship display: Some effects of third-party generosity on audience perceptions". In: *British Journal of Psychology* 101.1, pp. 123–135.

Boehm, Robert and Tobias Regner (Oct. 2013). "Charitable giving among females and males: an empirical test of the competitive altruism hypothesis". In: *Journal of Bioeconomics* 15.3, pp. 251–267.

Van Vugt, Mark and Wendy Iredale (Feb. 2013). "Men behaving nicely: Public goods as peacock tails". In: *British Journal of Psychology* 104.1, pp. 3–13.

Ottoni-Wilhelm, Mark, Lise Vesterlund, and Huan Xie (Sept. 2014). *Why Do People Give? Testing Pure and Impure Altruism*. Working Paper 20497. National Bureau of Economic Research.

Charness, Gary, Uri Gneezy, and Austin Henderson (May 2018). "Experimental methods: Measuring effort in economics experiments". In: *Journal of Economic Behavior & Organization* 149, pp. 74–87.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn (Jan. 2012). "Experimental methods: Between-subject and within-subject design". In: *Journal of Economic Behavior & Organization* 81.1, pp. 1–8.

Chen, Daniel L., Martin Schonger, and Chris Wickens (Mar. 2016a). "oTree – An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (Oct. 2014). "hroot: Hamburg Registration and Organization Online Tool". In: *European Economic Review* 71, pp. 117–120.

Meyer, Christian Johannes and Egon Tripodi (2018). *Image Concerns in Pledges to Give Blood: Evidence from a Field Experiment*. SSRN Working Paper 3132289.

Paul-Ehrlich-Institut (2018). *Preliminary Report for 2018*. Report on notifications pursuant to Section 21 TFG (German Transfusion Act). Langen: Paul-Ehrlich-Institut.

Trimborn, Marion (May 2009). "Rohstoff Blut: 1000 Euro für fünf Liter Blut". In: *Frankfurter Allgemeine Zeitung*.

Toner, Richard W. et al. (Aug. 2012). "Costs to hospitals of acquiring and processing blood in the US". In: *Applied Health Economics and Health Policy* 9.1, pp. 29–37.

World Health Organization (2017). *Global Status Report on Blood Safety and Availability 2016*. Tech. rep. Geneva: World Health Organization.

Whitaker, Barbee et al. (June 2016). *Trends in United States blood collection and transfusion: results from the 2013 AABB Blood Collection, Utilization, and Patient Blood Management Survey*. Tech. rep.

Greinacher, Andreas et al. (Apr. 2011). "Implications of demographics on future blood supply: a population-based cross-sectional study." In: *Transfusion* 51.4, pp. 702–709.

Offergeld, R. et al. (Feb. 2005). "Human immunodeficiency virus, hepatitis C and hepatitis B infections among blood donors in Germany 2000-2002: risk of virus transmission and the impact of nucleic acid amplification testing". In: *Euro Surveillance* 10.2, pp. 8–11.

Toner, Richard W. et al. (2011). "Costs to hospitals of acquiring and processing blood in the US: a survey of hospital-based blood banks and transfusion services". In: *Applied Health Economics and Health Policy* 9.1, pp. 29–37.

Baetschmann, Gregori, Kevin E Staub, and Rainer Winkelmann (2015). "Consistent estimation of the fixed effects ordered logit model". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.3, pp. 685–703.

Chamberlain, Gary (1980). "Analysis of Covariance with Qualitative Data". In: *The Review of Economic Studies* 47.1, pp. 225–238.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018). "Measuring and bounding experimenter demand". In: *American Economic Review* 108.11, pp. 3266–3302.

Harrison, Glenn W, Jimmy Martínez-Correa, and J Todd Swarthout (2014). "Eliciting subjective probabilities with binary lotteries". In: *Journal of Economic Behavior & Organization* 101, pp. 128–140.

Fehr, Ernst and Klaus M. Schmidt (Aug. 1999). "A Theory of Fairness, Competition, and Cooperation". In: *The Quarterly Journal of Economics* 114.3, pp. 817–868.