Destructive Communication*

Vasilisa Werner† Maximilian Andres‡

November 7, 2024

Abstract

This paper studies whether polarization offsets the well-documented positive effect of communication on trust. In an experimental setting, we vary the degree of polarization and the participants' ability to communicate. When participants are polarized enough, communication no longer improves trust and even harms trustworthiness. Using unsupervised machine learning, we document that a substantial fraction of individuals focus their communication on being polarized. This leads to a destructive effect of communication on both trust and trustworthiness.

JEL Codes: C92, D83, D91

Keywords: Communication; Polarization; Trust; Trustworthiness; Out-group discrimination; Online experiment

*We are grateful to Lisa Bruttel, Katharina Erhardt, Simon Martin, Hans-Theo Normann, Hannah Schildberg-Hörisch, Marco Schwarz, Bertil Tungodden, Georg Weizsäcker and the participants of the Bergen-Berlin Behavioral Economics Workshop, the DICE Brown Bag Seminar, IMEBESS 2024 and M-BEES 2024 for helpful comments and suggestions. We also thank Birte Prado Brand, Lennart Rehm, and Jonas Voigt for their excellent research assistance. Financial support from the Volkswagen Foundation is gratefully acknowledged. Experimental design, hypotheses, and analysis plan are preregistered via AEA RCT Registry: Andres, Maximilian and Vasilisa Werner. 2023. "Deterioration of Trust." AEA RCT Registry. September 15.

†Universität Potsdam and Berlin School of Economics, August-Bebel-Straße 89, 14482 Potsdam, email: vasilisa.werner@uni-potsdam.de

‡Universität Potsdam and Berlin School of Economics, August-Bebel-Straße 89, 14482 Potsdam, email: maximilian.andres@uni-potsdam.de

1 Introduction

Interpersonal communication is vital for social welfare by conveying information, resolving conflicts, negotiating agreements, building trust, fostering empathy, and promoting collaboration. It consistently improves the outcomes of social interactions: trust and trustworthiness, cooperation rates, public goods contributions, and the outcomes of other bargaining games.[1] The empirical evidence for the positive effect of communication is vast but *what if the communicators do not get along*? This paper challenges the universality of the positive effect of communication by focusing on individuals with polarized opinions about an important yet payoff-irrelevant topic.

Polarization manifests in numerous economic interactions, resulting in detrimental consequences for social welfare. Although polarization-driven discrimination occurs in a wide range of settings, it appears the strongest in the morality-based context,[2] e.g., for the political or religious views of the group members.[3] Balliet et al. (2018) demonstrate how political ideology influences cooperation in Prisoner's Dilemma, with both Republicans and Democrats displaying a lower tendency to cooperate outside of their respective groups. In this paper, we utilize the dissimilarity of views of pro-life and pro-choice individuals. Negative out-group effects often remain substantial even when the grouping is less meaningful, e.g., a preference for a Kandinsky or a Klee painting (O. Aksoy, 2019) or even not meaningful at all, e.g., group blue vs. group red (Charness, Di Bartolomeo, et al., 2024).[4]

In contrast to the previous literature, we show theoretically and experimentally that communication *harms* social welfare when polarization is sufficiently strong. In an online experiment, we vary whether subjects can communicate with each other and how polarized they are: from no polarization to strong polarization.[5] While no polarization implies that subjects do not receive any additional information about the person they interact with, strong polarization means a pro-choice individual knowingly interacts with a pro-life one. Based on theoretical predictions, we hypothesize that stronger polarization makes communication less beneficial for social welfare in economic interactions.

In stark contrast to the well-documented positive effect of communication, we find that

[1] Ben-Ner et al. (2011) and Charness and Dufwenberg (2006) document positive effects of communication for trust and trustworthiness. Andres (2023) and the literature therein show that communication improves cooperation rates in the Prisoner's dilemma setup. Communication further benefits public goods contributions (Dawes et al., 1977), and outcomes of the Dictator and Ultimatum games (Andreoni and Rao, 2011; Xiao and Houser, 2005).

[2] Morality-based out-group discrimination is a type of discrimination based on dissimilarity of beliefs and values. Common examples of morality-based groups are political and religious views (Grigoryan et al., 2023).

[3] See Parker and Janoff-Bulman (2013), Shu et al. (2012), and Weisel and Böhm (2015).

[4] Charness, Di Bartolomeo, et al. (2024) have shown that communication can overcome out-group discrimination about the assignment to an exogenous group with superficial meaning. Our treatments include a benchmark for groups without meaning as well and validate this finding in the Trust Game setup.

[5] We focus on out-group pairs, i.e., two subjects always belong to different groups.

when individuals are polarized, communication not only helps less but becomes destructive, i.e., harmful to the outcomes compared to no communication. Importantly, we document that a large share of subjects focus on their polarized opinions during communication, and, if that is the case, the outcomes of their interaction deteriorate even more drastically compared to no communication.

In this paper, we focus on trust and trustworthiness most commonly measured by the outcomes of the Trust Game (originally proposed by Berg et al., 1995). In the Trust Game, Player A (Trustor) sends a share of her endowment to Player B (Trustee). The amount sent gets multiplied and the trustee decides how much to send back. Contrary to theoretical predictions with selfish players, experimental evidence documents that participants display significant levels of trust and trustworthiness[6] measured by the amounts shared by the Trustor and shared back by the Trustee, respectively. In classic[7], multiplayer[8] and hidden-action[9] Trust Games, in the lab, the field and online,[10] trust and trustworthiness tend to benefit significantly from communication between players.

The novelty of this paper is three-fold. Firstly, the key contribution of our paper is providing a unique setup in which communication not only does not promote trust and trustworthiness but even harms these outcomes. To the best of our knowledge, this paper is the first to provide a counter-example to a well-established universality of the positive effect of interpersonal communication on the outcomes of social interactions.

Secondly, we provide detailed insights into the impact of the heterogeneous communication content. The content differs vastly and its impact on trust and trustworthiness varies from significantly positive if individuals choose to focus on the task to significantly negative when they discuss their differences of opinions.

Finally, our paper sheds light on out-group discrimination in general and morality-based out-group discrimination in particular. The evidence we document is mixed. With communication, both trust and trustworthiness are harmed by polarization induced by morality-based groups compared to groups with superficial meaning. However, without communication, morality-based discrimination does not occur to a significant extent for both trust and trustworthiness.

The rest of this paper is organized as follows. Section 2 describes the experimental design and summarizes technical details and implementation. Section 3 sets up the theoretical frame-

[6]See, e.g., B. Aksoy et al. (2018), Burks et al. (2003), Glaeser et al. (2000), McCabe, Rassenti, et al. (1998), and McCabe, Rigdon, et al. (2007).

[7]See, e.g., Babin and Chauhan (2023).

[8]See, e.g., Sheremeta and Zhang (2014).

[9]See Charness and Dufwenberg (2006), Ederer and Schneider (2022), Goeree and Zhang (2014), and Ismayilov and Potters (2016).

[10]See, e.g., Fiedler and Haruvy (2009).

work and presents our hypotheses. We provide a detailed discussion of experimental results in Section 4. Section 5 concludes.

2 Experimental design

In real-world situations, people tend to decide whether they trust others based on numerous aspects and over a longer period of time. In general and for economists in particular, these relevant characteristics are often hard to observe and classify. An experimental approach can thus be beneficial to induce and isolate out-group discrimination by reducing classification to only one observable characteristic.

General setup. Subjects are randomly matched in pairs to play a one-shot Trust Game. In the Trust Game, players are assigned one of two roles: Trustor (first mover) or Trustee (second mover). The Trustor decides how much of her endowment to share with the Trustee. The shared amount gets multiplied and transferred to the Trustee. The Trustee can then decide how to split the money between two players, i.e., how much to share back.[11] In our experimental treatments, we vary two aspects: subjects' ability to communicate with each other within pairs, and the group assignment. We summarize them in Table 1.

In treatments with communication (*Comm*), participants enter a chat where they can write to their matched partner freely for 180 seconds before they make decisions on how much they want to share and share back in the Trust Game. In treatments without communication (*NComm*), subjects cannot talk before making a decision.

Table 1: Experimental treatments.

Group assignment	Groups with meaning	Groups without meaning	No groups
Communication	*M/Comm*	*NM/Comm*	*Base/Comm*
No communication	*M/NComm*	*NM/NComm*	*Base/NComm*

Subjects can be a member of groups C or L and they choose their group membership and observe the group membership of their matched partner. We introduce a baseline treatment (*Base*) where subjects do not belong to any group. In treatments *NM*, groups C and L have no further meaning. In treatments *M*, we introduce meaning to the group assignment: We explain that C stands for pro-choice and L stands for pro-life and ask subjects to choose their group membership based on their beliefs. In treatments *M*, we provide the following explanation.

[11] We provide further details and the numerical calibration in our experiment below.

> Please indicate that you would like to belong to group C if you identify as Pro-Choice. According to the Cambridge Dictionary, pro-choice (adj.) means supporting the belief that a pregnant woman should have the freedom to choose an abortion if she does not want to have a baby.
>
> Please indicate that you would like to belong to group L if you identify as Pro-Life. According to the Cambridge Dictionary, pro-life (adj.) means opposed to the belief that a pregnant woman should have the freedom to choose an abortion if she does not want to have a baby.

We use the Cambridge Dictionary for these definitions not only because it is a reputable source but also because the difference in the wording of definitions of pro-life and pro-choice is minimal. We regretfully acknowledge the use of non-inclusive language in these definitions.

Grouping variable. We choose pro-life and pro-choice as a group variable in treatments *M* because it fulfills two major criteria. Firstly, in expectation, it creates a strong meaningful separation between two groups concerning a fundamental belief. Secondly, it naturally consists of only two categories (in contrast to, e.g., political views that, especially in the non-US context, might contain several options). To address the potentially mixed views on the topic, we elicit the intensity of participants' preference for the group membership on an 11-point Likert scale from 0 (completely indifferent) to 10 (extremely preferred) in addition to eliciting subjects' binary preference for group membership.

2.1 Timeline of the experiment.

In this section, we discuss our experiment step-by-step and summarize the timeline in Figure 1.

Instructions and screening. The online experiment starts with the attention checks. It allows us to screen out the subjects who do not pay attention to instructions. Subjects who passed the attention check read general instructions and explanations of the Trust Game.[12] After reading the instructions, subjects must answer a set of control questions to proceed.

Grouping and matching. In treatments *M*, we explain what the group membership in C and L stands for. Then, we elicit the binary choice for the group membership and the intensity of the preference for the group membership in treatments *M* and *NM*. Notably, subjects in *M* make an informed choice (aligned with their pro-choice or pro-life views) while those in *NM* choose between two letters with no further meaning.

[12] We provide the experimental instructions in Appendix C.

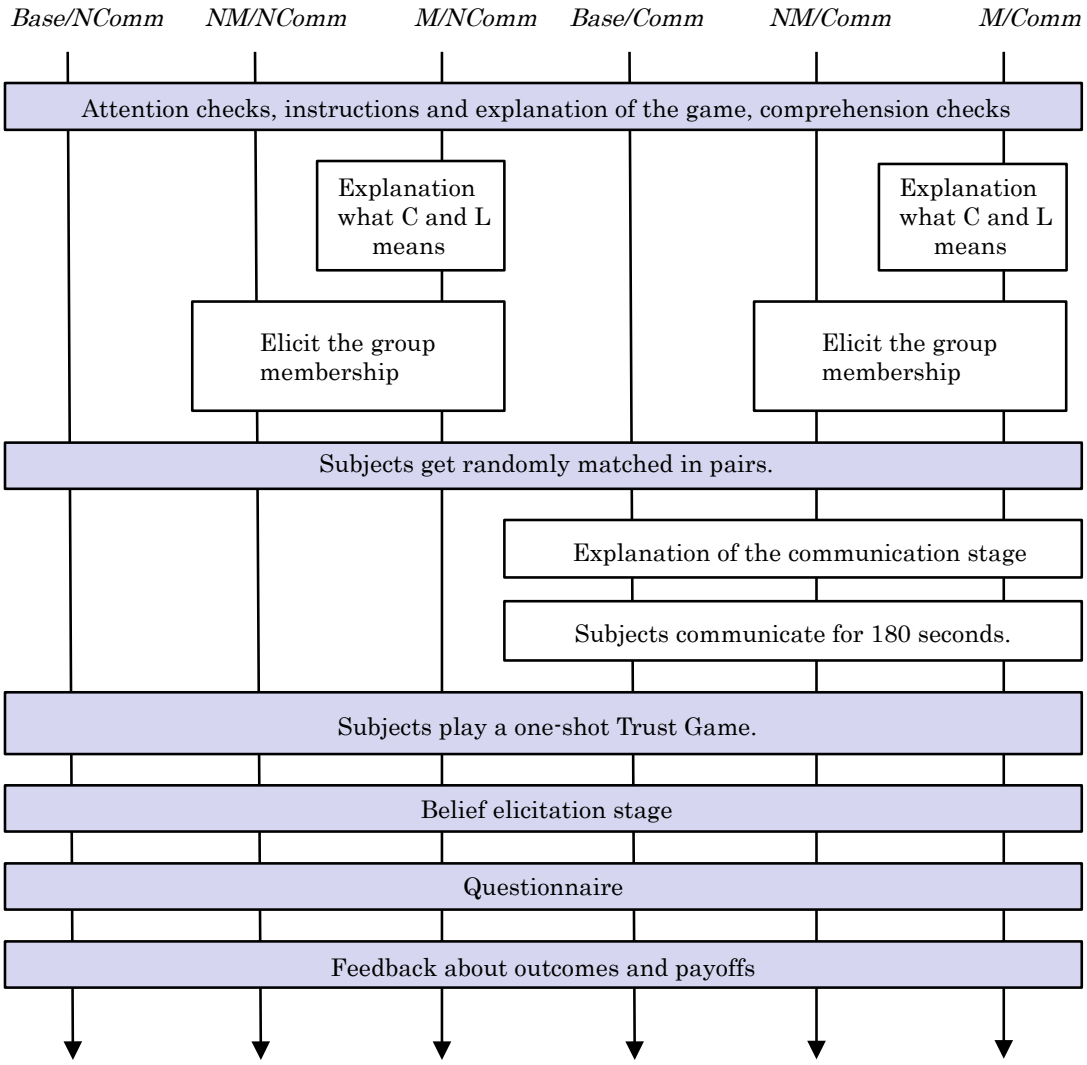


Figure 1: Timeline of the experiment

Note: This figure provides an overview of all stages of the experiment. The highlighted stages are the same for all treatments.

In the next step, we match subjects in pairs. As our study explicitly focuses on the out-group effects, we match subjects into pairs based on their random time of arrival to the matching stage and their group membership. It means that we always match one subject from group C to one subject from group L in treatments *NM* and *M*. Participants observe their partner's group membership. Importantly, matched subjects always belong to the same treatment. It means that there is no information asymmetry between them. Either both subjects in the pair know the meaning behind the group membership or neither knows.

Communication stage. In treatments *M/Comm*, *NM/Comm* and *Base/Comm*, we provide additional instructions about the communication stage. For 180 seconds, subjects can chat before they enter the decision stages. Their communication is open, i.e., they can type their messages and send them to their matched partner. There is no predetermined structure of their communication. Subjects cannot skip the communication stage. They also cannot shorten or extend the communication time. There are no restrictions on topics subjects can discuss and no priming or nudging suggesting they should discuss any particular topic. However, they are instructed not to share personal information such as names, phone numbers, etc.

Stage game. We endowed each participant with 10 points (ECU). Participant A can decide how many of these 10 points they want to share with Participant B. Participant B receives triple the amount of points that Participant A shared. Participant B can then decide to share their total amount of points (their endowment plus the share from Participant A) equally with Participant A or keep all points to themselves. Importantly, sharing equally means that Participant B splits points such that both participants receive the same amount of points.

We use the strategy method to elicit subjects' decisions. First, we asked subjects to pick the amount they would share, given they were Participant A. Second, for each amount shared (from 0 to 10), they decided whether they would share the points equally or keep all the points to themselves, given they were Participant B. This method allows us to elicit a switching point (from keeping to sharing) for each subject. Finally, a random mechanism decides whether they are Participant A or B with a 50% chance.

Belief elicitation. After subjects have made their decision and before they receive feedback, we elicit their beliefs about their partner's choices in both roles, i.e., the shared and the shared back amounts. We ask subjects for their best assessment of "the number of points (0-10) that the person you are paired with decided to share in the role of Participant A" as well as of "the minimal number of points (0-10) that the person you are paired with decided to receive to share all the points equally in the role of Participant B". We incentivize belief elicitation using the binarized scoring rule (Hossain and Okui, 2013) without explaining the details of monetary incentives to subjects (Danz et al., 2022).[13] Following Enke and Graeber (2023), we elicit subjects' cognitive uncertainty by asking subjects how confident they are in their beliefs.

Questionnaire. In the post-experimental survey, we collect data on subjects' socioeconomic characteristics, including age, gender, and education level. We further elicit whether individuals

[13]The detailed calculations were available to participants at the end of the experiment.

identify as pro-choice or pro-life in treatments *NM* and *Base*. Moreover, we use validated survey questions to elicit risk aversion following Falk et al. (2023), loss aversion (Fehr and Goette, 2007; Gächter et al., 2022) and social image concerns (Ewers and Zimmermann, 2015; Petrishcheva et al., 2023). After the questionnaire is complete, subjects observe the outcome of the Trust Game and the respective payment, as well as the payments for belief elicitations and the incentivized items from the post-experimental survey.

Technical details and procedure. We preregistered our experimental design, hypotheses, and analysis plan at the AEA Registry.[14] We conducted our experiment online on Prolific in September and October 2023. We collected data from 306 United States residents. Table 2 summarizes the number of independent observations per treatment. For treatments without communication, each subject represents one independent observation. For treatments with communication, one independent observation is a matched pair of subjects. In line with our preregistration, we oversampled treatment *M/Comm* to allow for heterogeneity analysis for different communication clusters.

Table 2: Number of independent observations per treatment.

	M	*NM*	*Base*	Total
Communication	72	24	24	120
No communication	22	20	24	66
Total	94	44	48	186

We conducted attention and comprehension checks in line with Prolific guidelines. There were two attention checks in our study. Subjects could not participate in our experiment if they answered both attention checks incorrectly. Furthermore, there were four comprehension questions to ensure participants understood the instructions. Each participant had two chances to answer each comprehension question correctly. After two unsuccessful attempts (in the same question), they could not continue the study. Participants could message us through Prolific at any time if they had questions.

Balancedness. Women constitute 48% of our sample. In each treatment, there are at least 47% and at most 50% of female participants. The differences in gender composition are not statistically significant between treatments (two-sided MWU tests, $p = 0.764$, $p = 1.000$ and $p = 0.874$ for *M/Comm* vs. *M/NComm*, *NM/Comm* vs. *NM/NComm* and *Base/Comm* vs. *Base/NComm*, respectively). Participants are 40 years old on average and the average age is

[14] Andres, Maximilian and Vasilisa Werner. 2023. "Deterioration of Trust." AEA RCT Registry. September 15. https://doi.org/10.1257/rct.10064-1.0

not statistically significantly different between treatments (two-sided MWU tests, $p = 0.788$, $p = 0.973$ and $p = 0.139$ for *M/Comm* vs. *M/NComm*, *NM/Comm* vs. *NM/NComm* and *Base/Comm* vs. *Base/NComm*, respectively).

Manipulation check. In addition to eliciting participants' group membership in treatments *NM* and *M*, we also elicit the intensity of group membership preference on an 11-point Likert scale from 0 (completely indifferent) to 10 (extremely preferred). We summarize the average intensity per treatment in Table 3.

Table 3: Summary statistics: Intensity of group membership preference per treatment.

Group assignment	*M*	*NM*
Communication	7.88	3.08
	(1.68)	(2.07)
No communication	7.77	3.35
	(2.31)	(2.93)

Note: Standard deviations are in parentheses. The intensity of group membership preference is on an 11-point Likert scale from 0 (completely indifferent) to 10 (extremely preferred).

In treatments *M*, where subjects report whether they identify as pro-life or pro-choice, the intensity of their preference for group membership is 7.88 and 7.77 in *Comm* and *NComm*, respectively. In treatments *NM*, where subjects report whether they want to be a member of groups C or L with no further meaning, their preference is, unsurprisingly, significantly weaker (one-sided MWU tests, $p < 0.001$ for *M/Comm* > *NM/Comm* and *M/NComm* > *NM/NComm*). The average intensity is 3.08 in *NM/Comm* and 3.35 in *NM/NComm*. The intensity of group membership preference is not significantly different for a given group assignment type (two-sided MWU tests, $p = 0.722$ for *M/Comm* vs. *M/NComm* and $p = 0.952$ for *NM/Comm* vs. *NM/NComm*).

3 Theoretical framework and hypotheses

To systematically address our research questions, Section 3.1 establishes a theoretical framework. Section 3.2 presents hypotheses based on theoretical predictions.

3.1 Theoretical framework

In the trust game, two agents $i \in \{A, B\}$ interact with each other. Each agent receives an initial endowment of ten tokens. We denote this strategy by $S \in \{0, 1, ..., 10\}$. Following agent A's decision, agent B receives the tripled amount of tokens $3 \cdot S$. Now, agent B must decide how many tokens she sends back to agent A: $R = k \cdot (3 \cdot S)$ where $k : \{0, 3, ..., 30\} \rightarrow \{0, 1, ..., 30\}$

and $k \in [0, 1]$. These strategies result in the monetary payoffs of $\pi_A(S, R) = 10 - S + R$ and $\pi_B(S, R) = 10 + 3 \cdot S - R$ for agents A and B, respectively.

A selfish agent B always prefers to keep all the tokens sent by agent A as $\frac{\partial \pi_B(S,R)}{\partial R} < 0$. Anticipating this behavior, a rational agent A should send nothing because $\frac{\partial \pi_B(S,R=0)}{\partial S} < 0$. This behavior, however, is in stark contrast to the empirical data, suggesting that individuals send positive monetary transfers in the Trust Game. Conceptually, the agent A's trust only pays off if agent B has other-regarding preferences (see Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fehr and Schmidt, 1999). Thus, it might be agent A's anticipation of the other-regarding preferences of agent B resulting in the observed gap between the empirical data and the selfish behavior. These positive transfers imply the interior solution i.e., a positive amount sent and returned. Therefore, closely following Fiedler, Haruvy, and Li (2011) who build on the seminal work of Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002), we incorporate other-regarding preferences to the agent B's utility function. We represent the utility function of agent B as follows:

$$U_B(\pi_A(S, R), \pi_B(S, R), \beta) = \pi_B(S, R) - \beta \cdot \left(\frac{\pi_B(S, R) - \pi_A(S, R)}{\pi_A(S, R) + \pi_B(S, R)} \right)^2, \tag{1}$$

where $\beta > \frac{40}{3}$ is the inequality aversion parameter.[15] We assume the second term of Equation (1) to be quadratic and relative to "total size of the pie" (i.e., $\pi_A(S, R) + \pi_B(S, R)$) to allow for interior solutions. The best response function of agent B is the following:[16]

$$R^*(S, \beta) = 2 \cdot S - \frac{(10 + S)^2}{2 \cdot \beta}. \tag{2}$$

Agent A maximizes her utility function conditional on the agent B's best response:

$$U_A(\pi_A(S, R = R^*(S, \beta)) = \pi_A(S, R = R^*(S, \beta)). \tag{3}$$

Agent A gives optimally shares[17]

$$S^*(\beta) = \beta - 10. \tag{4}$$

Hence, given agent A's best response, the optimal amount sent back is[18]

$$R^*(S = S^*(\beta), \beta) = \frac{3}{2} \cdot \beta - 20. \tag{5}$$

[15] For $\beta > \frac{40}{3}$, which we assume throughout this paper, we obtain an interior solution in the trust game. See Appendix A.4 for a proof.

[16] We provide all details related to Equation 2 in Appendix A.1.

[17] See Appendix A.2 for details related to Equation 4.

[18] See Appendix A.3 for details related to Equation 5.

Lemma 1. *Agents with stronger other-regarding preferences share and share back more, i.e.,* $\frac{\partial S^*(\beta)}{\partial \beta} > 0$ *and* $\frac{\partial R^*(S=S^*(\beta),\beta)}{\partial \beta} > 0$.

Proof. See Appendix A.5 for proof. □

Intuitively, both $S^*(\beta)$ and $R^*(S = S^*(\beta), \beta)$ increase in the strength of the inequality aversion. A higher (lower) value of β implies agent B gains more (less) disutility from inequality between agents. To maximize her utility, agent B will send back a certain positive amount to minimize the disutility from inequality. Anticipating this behavior, agent A will send the more (higher S^*) the more inequality-averse agent B is. Thus, this theoretical framework captures the behavior observed in the empirical data.

Communication and Social Distance. First, let us introduce the social distance parameter $d \geq 0$, where $d = 0$ corresponds to playing with an equivalent agent and increasing the distance denotes larger social differences between agents. Intuitively, having conflicting opinions increases the social distance between agents. Then, $d = 0$ implies no conflict of opinions, and increasing d corresponds to a stronger conflict of opinions.[19]

Second, we introduce the communication parameter $c \in \{0, 1\}$, where $c = 0$ denotes the absence of communication and $c = 1$ denotes the presence thereof. We formulate three crucial assumptions about the structure of other-regarding preferences.

Assumption 1. *Other-regarding preferences decrease in the distance d:* $\frac{\partial \beta}{\partial d} < 0$.

The larger the social distance is, the less agents care about the inequality between them. This implies that more polarized agents have weaker other-regarding preferences towards each other.

Assumption 2. *Interpersonal communication strengthens other-regarding preferences:* $\beta_{c=1}(d = 0) > \beta_{c=0}(d = 0)$.

This assumption captures the well-documented positive effect of communication on the outcomes of social interactions. Agents who have an opportunity to communicate with each other care more about the inequality between them.

Assumption 3. *Other-regarding preferences decrease in the distance d faster with communication than without communication:* $\frac{\partial \beta_{c=0}}{\partial d} > \frac{\partial \beta_{c=1}}{\partial d}$.

[19] Our theoretical framework allows for a natural extension incorporating in-group favoritism, i.e., the social distance between two agents *shrinking* compared to the baseline.

This assumption establishes how communication and social distance interact. Essentially, we assume that communication impact on agents' other-regarding preferences is the smaller, the more polarized (i.e., distant) agents are. Next, we derive hypotheses based on this theoretical framework.

3.2 Hypotheses

For the Hypotheses below, we refer to the "out-group effects". We assume that out-group effects are absent in treatments *Base* (by definition, i.e., no groups) and present in treatments *NM* and *M*, with out-group effects being stronger in *M* than in *NM* due to the polarizing nature of group membership. These out-group effects correspond to distance d, namely, $d_M > d_{NM} > d_{Base} = 0$.

First, we discuss our predictions about trust and trustworthiness differences between treatments *M*, *NM*, and *Base*. In Hypothesis 1, we compare how subjects share and share back in the Trust Game if their group membership becomes more polarizing keeping the communication (or lack thereof) constant. Hypothesis 1 postulates that when subjects belong to different groups increasing differences will harm trust and trustworthiness.

Hypothesis 1. (Shared and shared back amounts) *The amounts shared and shared back in the Trust Game decrease if the out-group effects become stronger (i.e., distance d increases).*

Formally, Hypothesis 1 aims to test whether $\frac{\partial S}{\partial d} < 0$ and $\frac{\partial R}{\partial d} < 0$. According to our theoretical predictions, $\frac{\partial S}{\partial \beta} \cdot \frac{\partial \beta}{\partial d} < 0$ because $\frac{\partial S}{\partial \beta} > 0$ due to Lemma 1 and $\frac{\partial \beta}{\partial d} < 0$ due to Assumption 1. Likewise, according to our theoretical predictions, $\frac{\partial R}{\partial \beta} \cdot \frac{\partial \beta}{\partial d} < 0$ because $\frac{\partial R}{\partial \beta} > 0$ due to Lemma 1 and $\frac{\partial \beta}{\partial d} < 0$ due to Assumption 1.

We hypothesize that the amount sent as Participant A is, on average, larger in *Base* than in *NM*, and larger in *NM* than in *M*. Concerning the amount shared back, for a strategy method setup, it implies that the minimal amount received to share back as Participant B is smaller in *Base* than in *NM*, and smaller in *NM* than in *M*.

Hypothesis 2. (Communication quality) *Communication becomes unfriendlier if the out-group effects become stronger (i.e., distance d increases).*

Communication quality is a direct prerequisite for Assumption 3 stating that the shift in social preferences due to communication decreases in the distance d. Conceptually, we attribute this shift to the quality of communication, i.e., how productive and/or kind communication is. Thus, the evidence in favor of Hypothesis 2 justifies Assumption 3, i.e., $\frac{\partial \beta_{c=0}}{\partial d} > \frac{\partial \beta_{c=1}}{\partial d}$.

The next hypothesis summarizes our predictions about whether communication boosts trust and trustworthiness. There is mounting experimental evidence that communication possibilities

improve both trust and trustworthiness (see Ben-Ner et al., 2011, and the literature therein). Our paper focuses on increasing differences between subjects which potentially lead to a higher degree of hostility in certain environments, for example, in treatment *M/Comm*.

Hypothesis 3. (Communication impact) *The positive effect of communication on sharing and sharing back in the Trust Game becomes smaller if the out-group effects become stronger (i.e., distance d increases).*

To formalize Hypothesis 3, we focus on

$$\frac{\partial(S_{c=1} - S_{c=0})}{\partial d} = \frac{\partial S_{c=1}}{\partial d} - \frac{\partial S_{c=0}}{\partial d} = \frac{\partial S_{c=1}}{\partial \beta_{c=1}} \cdot \frac{\partial \beta_{c=1}}{\partial d} - \frac{\partial S_{c=0}}{\partial \beta_{c=0}} \cdot \frac{\partial \beta_{c=0}}{\partial d}. \tag{6}$$

According to Equation 4, $\frac{\partial S_{c=1}}{\partial \beta_{c=1}} = \frac{\partial S_{c=0}}{\partial \beta_{c=0}} = 1$, we simplify Equation 6 as follows:

$$\frac{\partial(S_{c=1} - S_{c=0})}{\partial d} = \frac{\partial \beta_{c=1}}{\partial d} - \frac{\partial \beta_{c=0}}{\partial d} < 0, \tag{7}$$

which holds according to Assumption 3. Analogously,

$$\frac{\partial(R_{c=1} - R_{c=0})}{\partial d} < 0. \tag{8}$$

Hence, we hypothesize that the effect of communication on shared amounts in the Trust Game is larger in *Base* than in *NM*, and larger in *NM* than in *M*. Likewise, we hypothesize that the effect of communication on shared amounts in the Trust Game is larger in *Base* than in *NM*, and larger in *NM* than in *M*. Our paper focuses on increasing differences between subjects and will potentially lead to higher degrees of hostility in certain environments, for example, if subjects belong to different groups in treatment *M/Comm*.

4 Results

This section is organized as follows. First, we report levels of trust and trustworthiness across treatments in Section 4.1. Then, we analyze communication content in treatments *M/Comm*, *NM/Comm*, and *Base/Comm* in Section 4.2. We discuss the impact of communication on sharing and sharing back in the Trust Game if the differences between subjects increase, i.e., the out-group effects get stronger in Section 4.3. Finally, we focus on beliefs in Section 4.4 and provide additional insights in Section 4.5. Throughout this section, we report p-values of one-sided[20] Mann-Whitney U tests with continuity correction unless specified otherwise.

[20]Following our preregistered analysis plan, we will rely on one-sided Mann-Whitney U tests due to the directional nature of our hypotheses.

4.1 Trust and trustworthiness.

First, we focus on the average treatment effects. We summarize the results on trust and trustworthiness in turn.

Trust. Figure 2 shows three crucial patterns in levels of trust between treatments. First, participants trust less as out-group effects get stronger when they can communicate with each other. In contrast to only 5.27 out of 10 ECU in treatment *M/Comm*, they share on average 6.92 and 7.02 out of 10 ECU in treatments *NM/Comm* and *Base/Comm*, respectively. While introducing groups per se does not lead to a significant decline in trust (*NM/Comm* vs. *Base/Comm*, $p = 0.409$), meaningful out-group effects result in a highly significant drop in trust levels compared to groups without meaning (*M/Comm* vs. *NM/Comm*, $p = 0.006$) and the baseline (*M/Comm* vs. *Base/Comm*, $p = 0.004$).

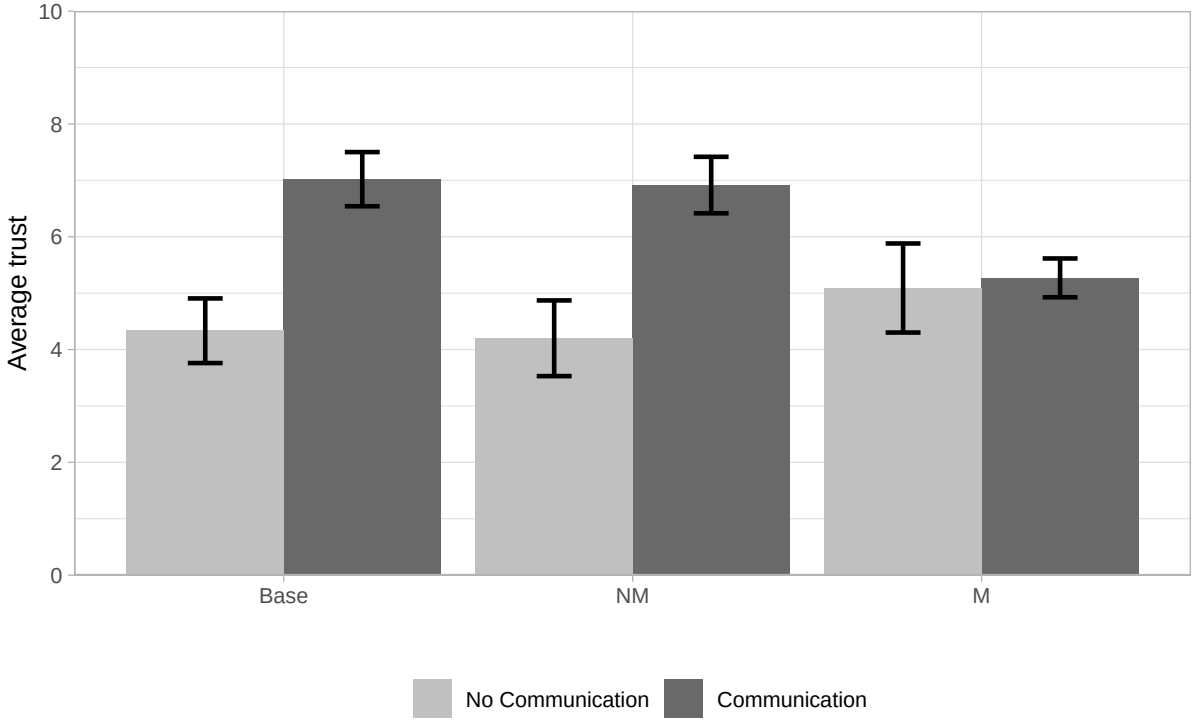


Figure 2: Average levels of trust per treatment. Error bars indicate standard errors.

Second, there are no significant differences in trust without communication. In *Base/NComm*, participants send, on average, 4.33 out of 10 ECU in the trust game. Introducing out-group without and with meaning has no significant effect on trust. Participants share 4.20 and 5.09 out of 10 ECU in treatments *NM/NComm* and *M/NComm*, respectively, and all the differences are not statistically significant.[21]

[21]The MWU test p-values are 0.767, 0.670, and 0.424 for comparing the levels of trust in *M/NComm* vs. *NM/NComm*, *M/NComm* vs. *Base/NComm*, and *NM/NComm* vs. *Base/NComm*, respectively.

The third insight in Figure 2 is that the positive effect of communication is fully offset by polarizing out-group effects. We compare treatments with and without communication for a given group assignment. In the baseline, communication has a strong and highly significant effect on trust, bringing the amounts participants shared from 4.33 to 7.02 out of 10 ECU (*Base/NComm* vs. *Base/Comm*, $p < 0.001$). Likewise, if subjects were in groups that have no further meaning, communication has shifted the shared amounts from 4.20 to 6.92 out of 10 ECU on average (*NM/NComm* vs. *NM/Comm*, $p = 0.001$). However, having stronger out-group effects erases the well-documented positive effect of communication. Subjects share on average 5.09 and 5.27 out of 10 ECU in treatments *M/NComm* and *M/Comm*, respectively, and the impact of communication on trust is not statistically significant ($p = 0.322$). Additionally, we conduct two one-sided t-tests comparing the levels of trust in *M/Comm* and *M/NComm* to document the null result ($p = 0.582$ and $p = 0.418$ for *M/Comm* $<$ *M/NComm* and *M/Comm* $>$ *M/NComm*, respectively). We have a detailed discussion of the impact of communication on trust in Section 4.3.

Trustworthiness. With a strategy method, individuals make a set of 11 binary decisions resulting in a single switching point. This switching point captures the minimal amount their matched partner has to share in the role of Player A for an individual to share back equally as Player B. Crucially, a lower switching point indicates a higher level of trustworthiness. Hence, to make the interpretation more intuitive, we revert the scale and measure trustworthiness as (10−switching point). This way, a higher number corresponds to more trustworthiness.

We summarize the average trustworthiness in each treatment in Figure 3. In treatments with communication, we observe a decline in trustworthiness as out-group effects get stronger. In *Base/Comm*, participants' trustworthiness is 7.48 on average, 7.15 in *NM/Comm* and only 6.39 in *M/Comm*. While the direct differences between treatments are statistically insignificant, with $p = 0.144$ between treatments *M/Comm* and *NM/Comm* and $p = 0.260$ between treatments *NM/Comm* and *Base/Comm*, the overall drop in trustworthiness between *M/Comm* and *Base/Comm* is statistically significant with $p = 0.048$.

Without communication, polarization does not have a significant effect on trustworthiness. Trustworthiness levels are 7.13, 7.40 and 7.41 in treatments *Base/NComm*,*NM/NComm* and *M/NComm*, respectively, and all the differences are statistically insignificant.[22] That is, we document that polarization per se does not significantly affect trustworthiness, only in combination with communication.

[22]The MWU test p-values are 0.642, 0.791, and 0.645 for comparing the levels of trustworthiness in *M/NComm* vs. *NM/NComm*, *M/NComm* vs. *Base/NComm*, and *NM/NComm* vs. *Base/NComm*, respectively.

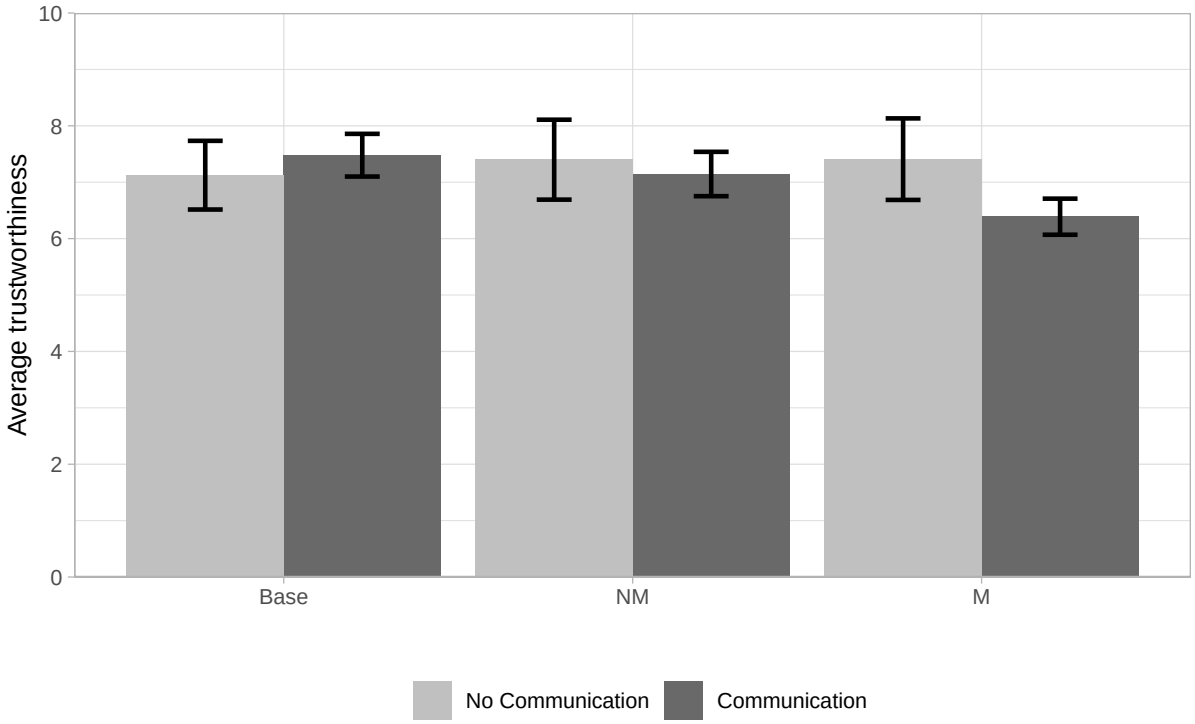


Figure 3: Average levels of trustworthiness per treatment. Error bars indicate standard errors.

The communication effect on trustworthiness is quite striking. First and foremost, communication does not improve trustworthiness in the baseline treatment (*Base/NComm* vs. *Base/Comm*, $p = 0.533$). Once the out-group effect is introduced in *NM*, we find suggestive evidence for a negative effect of communication on trustworthiness (*NM/NComm* vs. *NMComm*, $p = 0.102$). For meaningful polarizing out-group effect in *M*, communication has a strong and statistically significant negative effect (*M/NComm* vs. *M/Comm*, $p = 0.011$). We have a further discussion of the impact of communication on trustworthiness in Section 4.3.

4.2 Communication content.

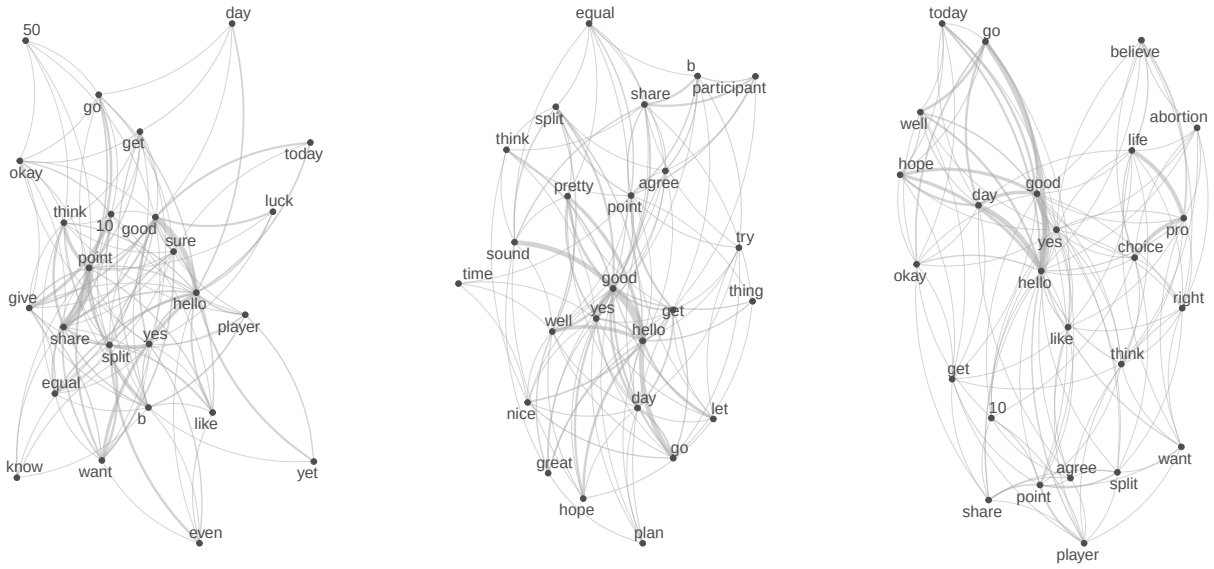
To analyze the content of communication, we use natural language processing. Natural language processing is a common approach to quantifying text data (see Andres, Bruttel, and Friedrichsen, 2023, and the literature therein). The main advantage of this approach is an objective estimation of the text data, contrary to the chats being manually reviewed by the researcher.

Before we analyze the content of all chats, we process the communication data, or corpus, to be expressive. We correct spelling mistakes, eliminate words without meaningful content, i.e., stopwords, such as "the", and reduce words to their linguistic root, such as "prefers" becomes "prefer". The corpus is a matrix where each row represents a chat or document, and each column represents a token. Tokens can be words in the documents, or, for example, a number. We

describe all technical details of communication analysis in Appendix B.

4.2.1 Communication content per treatment.

We provide a descriptive overview of communication content in Figure 4. It shows the 25 most frequent tokens per treatment and their corresponding phi-coefficient using connection lines to show the key distinguishing features between clusters. The phi-coefficient records the co-occurrences of tokens in each treatment. Thus, these figures show the key distinguishing features of the communication content between treatments.



(a) *Base/Comm* (b) *NM/Comm* (c) *M/Comm*

Figure 4: 25 most frequent tokens per treatment and their corresponding phi-coefficient by the lines between the tokens.

At first glance, communication in *Base/Comm* and *NM/Comm* seems rather similar featuring mostly discussion of the game ("split", "share", "equal"), different forms of agreement ("yes", "ok") and exchanging pleasantries (e.g., "good luck" and wishing each other to have a "nice day").

However, the most commonly used words in treatment *M/Comm* make a different first impression. In addition to the tokens mentioned above, there is a distinct pattern of individuals talking about "pro", "life" and "choice" as well as using words like "abortion". It indicates that many participants focus their communication on the polarizing topic, whereas this topic is absent in the other treatments.

To document the shift in communication content in a more structured way, we cluster the communication content into topics such that we can compare the shift of the polarizing topics across treatments.[23]

[23]We have explicitly pre-registered clustering data based on the communication content. Our pre-registered sample size accounts for clustering.

4.2.2 Hierarchical cluster analysis.

After transforming the chat messages into the matrix, we analyze the communication content using a Hierarchical Cluster Analysis, a common unsupervised machine learning method to cluster documents (see Andres and Bruttel, 2024, and the literature therein). To apply the Hierarchical Cluster Analysis, we use an agglomerative algorithm. This algorithm starts by treating each observation as a separate cluster and repeatedly extends the clusters until only one cluster exists. At each iteration, the two least distant clusters form a new one measuring the distance between clusters using their Euclidean distance. We run the algorithm on a binary dissimilarity matrix using the term-frequency-inverse-document-frequency of tokens that are not more sparse than 90% to consider only the most important tokens. Based on the mean silhouette width presented and explained in Appendix B.3, we determine the optimal number of clusters to be four.

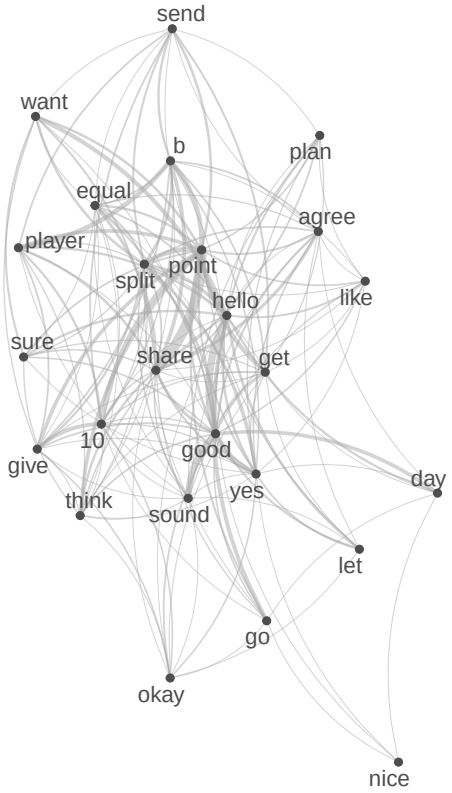
We summarize the result of the Hierarchical Cluster Analysis in Figure 5. It illustrates the 25 most frequent tokens in each cluster across treatments and their phi-coefficient to show the key distinguishing features between clusters. There are four distinct clusters. The cluster in Figure 5a contains tokens like "pro", "life", "choice", "abortion", "fetus", etc. These tokens indicate that this cluster mainly contains explicit talk about the out-group characteristic. We henceforth will refer to this cluster as *out-group*. Chats belonging to the cluster in Figure 5b contain attempts to coordinate the strategies represented by tokens like "share", "split", "equally", and "agree". Therefore, as subjects focus on their task, we refer to this cluster as *task*. The cluster in Figure 5c represents *small talk*. Subjects talk about the "weather", discuss what their plans for "today" are, and wonder whether their matched partner watched a football "game" last night. Finally, the cluster in Figure 5d represents subjects that abstain from communication almost entirely. They greet each other and then remain silent for the remainder of the 3-minute time frame. We refer to this cluster as *no interaction*.[24]

4.2.3 Polarization analysis.

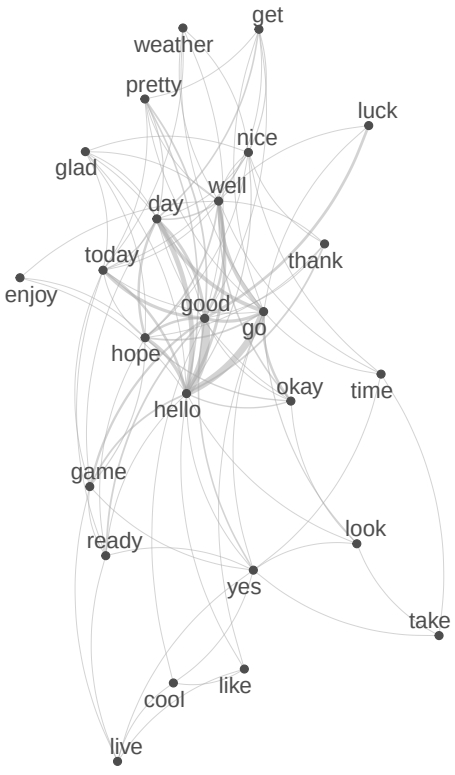
We classify communication clusters as follows. We refer to communication as being more polarized if the share of chats belonging to the *out-group* cluster increases and more productive if the share of chats in the *task* cluster increases. We view *small talk* and *no interaction* clusters as neutral. Table 4 displays the number of observations in each treatment split up by clusters.

[24]Interestingly, while participants exchange only 3.44 tokens on average in *no interaction* cluster, the average length of chats in other clusters is substantially longer. Subjects' average communication length is 27.24 tokens in *out-group*, 32.19 tokens in *task*, and 19.27 tokens in *small talk* clusters. The number of tokens is significantly lower in the *no interaction* cluster than in the *out-group* ($p < 0.001$), *task* ($p < 0.001$), and *small talk* ($p < 0.001$) cluster. The result looks similar once we compare the average length of chats between communication clusters across treatments.

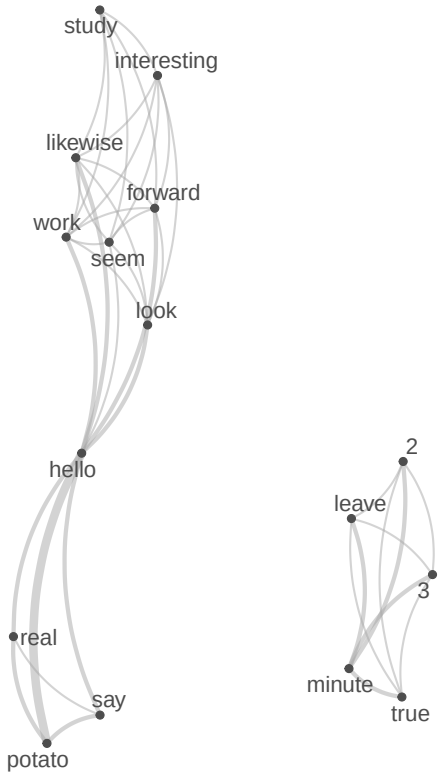
(a) *Out-group* cluster.



(b) *Task* cluster.



(c) *Small talk* cluster.



(d) *No interaction* cluster.

Figure 5: 25 most frequent tokens per communication cluster and their corresponding phi-coefficient illustrated by the lines between the tokens.

Table 4: Descriptive statistics for different communication clusters.

Treatment	Cluster	Share of obs.	Trust		Trustworthiness	
			Mean	Std. Dev.	Mean	Std. Dev.
M/Comm	*out-group*	23.6%	3.21	1.74	5.59	2.88
	task	27.8%	7.75	3.40	6.38	2.83
	small talk	38.9%	5.20	1.60	6.86	2.65
	no interaction	9.7%	3.50	2.89	6.50	2.29
NM/Comm	*out-group*	-	-	-	-	-
	task	62.5%	7.83	2.10	7.57	1.74
	small talk	33.3%	5.56	2.43	6.44	2.56
	no interaction	4.2%	4.00	0.00	6.50	0.00
Base/Comm	*out-group*	-	-	-	-	-
	task	75.0%	7.53	2.36	7.69	1.90
	small talk	20.1%	5.70	1.75	7.00	1.84
	no interaction	4.2%	4.50	0.00	6.00	0.00

Both in treatments *Base/Comm* and *NM/Comm*, individuals do not communicate about their out-group differences at all.[25] The relative number of observations in the *out-group* cluster to the *task* cluster is therefore larger in *M/Comm* than in either *NM/Comm* and *Base/Comm*. The differences are statistically significant in a one-sided Fisher exact test for *M/Comm* versus *NM/Comm* ($p < 0.001$), and *M/Comm* versus *Base/Comm* ($p < 0.001$).

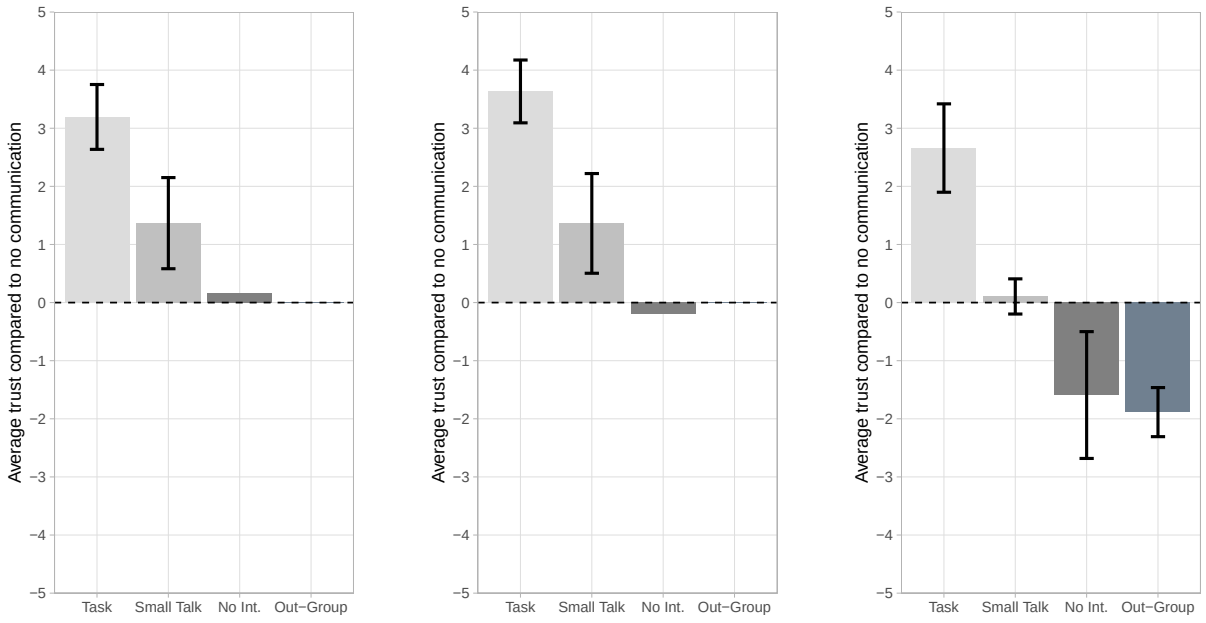
The fraction of groups in the *task* cluster accounts for 62.5% and 75.0% in treatments *NM/Comm* and *Base/Comm*, while it applies to only 27.8% of chats in treatment *M/Comm*. The difference in the fraction of groups in the *task* cluster relative to the *small talk* and *no interaction* is statistically significant in a one-sided Fisher exact test in *M/Comm* versus *NM/Comm* ($p = 0.029$), and *M/Comm* versus *Base/Comm* ($p = 0.002$). Interestingly, the differences between *NM/Comm* and *Base/Comm* are not statistically significant ($p = 0.267$). Thus, the data supports the hypothesis that communication is more polarized in *M/Comm* than in either in *NM/Comm* or *Base/Comm*. However, we detect no significant differences in communication structure between *NM/Comm* and *Base/Comm*. This result is mainly driven by the absence of *out-group* cluster in both treatments.

4.3 Communication impact on trust and trustworthiness.

In addition to the overview of communication clusters, Table 4 provides the average levels of trust and trustworthiness in each cluster by treatment.

Communication clusters and trust. Figure 6 illustrates the average trust per treatment by cluster. Importantly, it displays relative differences, i.e., deviations from the respective treat-

[25]The differences between *NM/Comm* and *Base/Comm* are hence not statistically significant ($p = 1$).



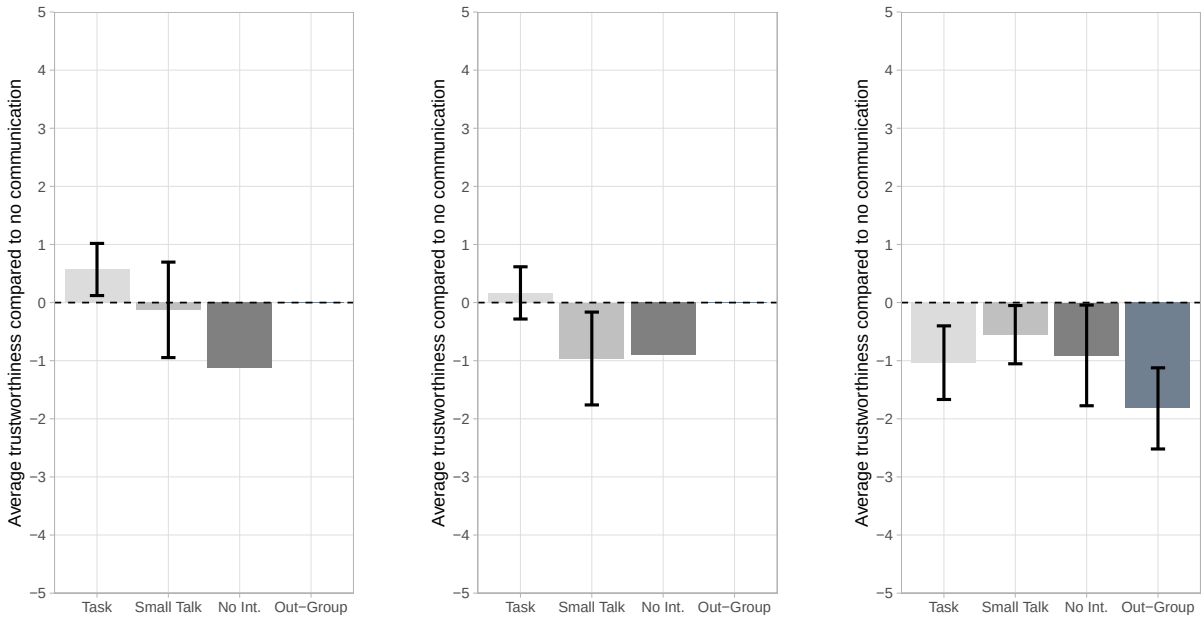
(a) *Base/Comm* (b) *NM/Comm* (c) *M/Comm*

Figure 6: Average trust per treatment by communication cluster compared to corresponding treatments without communication. Error bars indicate standard errors. Figures (a)-(c) display relative differences, i.e., zero in Figure (a) corresponds to the average level of trust in *Base/NComm*, zero in Figure (b) corresponds to the average level of trust in *NM/NComm*, and zero in Figure (c) corresponds to the average level of trust in *M/NComm*.

ments without communication. Focusing interpersonal communication on the task at hand boosts trust significantly compared to settings without communication possibilities. The average trust is significantly higher in groups discussing the task than in groups without communication: $p = 0.012$ in *M/Comm*, $p < 0.001$ in *NM/Comm* and $p < 0.001$ in *Base/Comm*. Thus, across treatments, our data supports the idea that productive communication and discussing trusting each other is associated with higher trust compared to settings without communication possibilities.

Small talk is associated with higher levels of trust compared to no communication too. However, the positive boost decreases in the polarization level. While the trust level in the *Small talk* cluster in *Base/Comm* is weakly significantly higher than in *Base/NComm* ($p = 0.076$) and in *NM/Comm* than in *NM/NComm* ($p < 0.100$), it is not significantly higher in *M/Comm* than in *M/NComm* ($p = 0.261$). The average trust level is not significantly higher in the *no interaction* cluster in *M/Comm* than in *M/NComm* ($p = 0.823$).

Counterproductive communication stands out in Figure 6c. The average trust level in groups with *out-group* communication is lower than in groups without communication ($p = 0.077$). Thus, when interpersonal communication focuses on the out-group identity, our data supports the hypothesis, that the average trust level can be higher without communication than with



(a) *Base/Comm* (b) *NM/Comm* (c) *M/Comm*

Figure 7: Average trustworthiness per treatment by communication cluster compared to corresponding treatments without communication. Error bars indicate standard errors. Figures (a)-(c) display relative differences, i.e., zero in Figure (a) corresponds to the average level of trustworthiness in *Base/NComm*, zero in Figure (b) corresponds to the average level of trustworthiness in *NM/NComm*, and zero in Figure (c) corresponds to the average level of trustworthiness in *M/NComm*.

communication.

Communication clusters and trustworthiness. Figure 7 illustrates the average trustworthiness per treatment by cluster compared to treatments without communication. The average trustworthiness is significantly lower in groups with communication than in groups without communication in *task* ($p = 0.048$), in *small talk* ($p = 0.061$) and in *no interaction* ($p = 0.086$) in treatment *M*. The average trustworthiness is marginally significantly lower in groups with communication than in groups without communication in *small talk* ($p = 0.075$) in treatment *NM*, but not in treatment *Base* ($p = 0.373$). Similarly, in *task* cluster, the average trustworthiness is not significantly lower in groups with communication than in groups without communication in treatment *NM* ($p = 0.239$) and in treatment *Base* ($p = 0.572$).

The average trustworthiness level in *M/Comm* is drastically lower in the *out-group* cluster compared to *M/NComm* indicating that individuals who engage in discussions regarding their differences display exceptionally low levels of trustworthiness ($p = 0.006$). Thus, across communication clusters, our data supports the idea that communication can harm trustworthiness if the grouping is polarizing enough.

4.4 Beliefs

Next, we study the beliefs across treatments and communication clusters summarized in Table 5. In addition to point beliefs, we elicit subjects' confidence in their assessment on a scale from 0 (not at all confident) to 10 (extremely confident). Point beliefs and confidence are jointly incentivized.

Table 5: Mean beliefs about trust and trustworthiness per treatment.

	Belief about trust		Belief about trustworthiness	
Treatment	Mean	Std. Dev.	Mean	Std. Dev.
M/NComm	3.82	2.54	5.59	2.58
M/Comm	5.28	2.96	4.50	2.39
NM/NComm	4.80	2.48	5.95	2.40
NM/Comm	6.65	2.56	3.81	2.06
Base/NComm	4.33	2.60	4.67	2.70
Base/Comm	7.18	2.52	3.41	2.45

Communication affects beliefs about sharing in the role of player A positively across treatments. The average belief about trust significantly improves in treatments with communication: $p = 0.025$ in *M*, $p = 0.011$ in *NM*, and $p < 0.001$ in *Base*. Average beliefs about trust are lower in *M/Comm* compared to the other treatments with communication – $p = 0.006$ compared to *Base/Comm* and $p = 0.023$ compared to *NM/Comm* – indicating that individuals believe their matched partner share less if their views are polarized.

For trustworthiness, communication has a negative effect on beliefs for sharing back in the role of player B. The average belief about trustworthiness in treatments with communication is significantly lower than in the treatments without communication in *M* ($p = 0.044$), in *NM* ($p < 0.001$) and in *Base* ($p = 0.062$). The beliefs about trustworthiness increase in *M/Comm* compared to the other treatments with communication. The belief of player A is significantly lower in *Base/Comm* than in *M/Comm* ($p = 0.020$). The result suggests the same direction of the effect in *NM/Comm* versus *M/Comm* with $p = 0.109$.[26]

In treatments with communication, subjects' confidence remains relatively stable. On average, they report 6.43, 6.38, and 6.41 for the amount shared in treatments *Base/Comm*, *NM/Comm*, and *M/Comm*, respectively. All differences between treatments are insignificant.[27] For the amount shared back, their confidence assessment is slightly higher in *Base/Comm* (6.96 on average) than in *NM/Comm* (6.21 on average, two-sided MWU test $p = 0.504$ for *Base/Comm*

[26] Beliefs about trust ($p = 0.256$) and trustwothiness ($p = 0.135$) are not significantly different in *Base/Comm* and *NM/Comm*.

[27] *Base/Comm* vs. *NM/Comm* with $p = 0.813$, *Base/Comm* vs. *M/Comm* with $p = 0.807$, and *M/Comm* vs. *NM/Comm* with $p = 0.869$ (two-sided MWU tests).

vs. *NM/Comm*) and *M/Comm* (6.00 on average, two-sided MWU test $p = 0.031$ for *Base/Comm* vs. *M/Comm*).

Without communication, beliefs are similar across treatments. The difference in beliefs about trust is insignificant in *Base/NComm* and *M/NComm* ($p = 0.261$) and weakly significant ($p = 0.090$) in *NM/NComm* in *M/NComm*. Likewise, the belief about trustworthiness are significantly different in *Base/NComm* and *M/NComm* ($p = 0.179$) as well as in *NM/NComm* and *M/NComm* ($p = 0.787$).

Additionally, in treatments without communication, we document a weakly significant decrease in confidence concerning the amount shared, that is, individuals are less sure about their assessment of their opponent's trust when they are more polarized. In *Base/NComm*, subjects' average assessment of confidence for the amount shared is 6.46 and it decreases to 5.27 and 5.00 in treatments *M/NComm* and *NM/NComm*, respectively.[28] For the amount shared back, subjects display relatively high levels of confidence in their assessment in *Base/Comm* (6.42 on average). Their confidence tends to be marginally lower in both *M/NComm* (5.77 on average) than in *NM/NComm* (4.80 on average).[29]

Belief accuracy. The accuracy of individual beliefs differs tremendously. The overall pattern of beliefs aligns relatively well with the pattern we observe in trust but not trustworthiness levels. For the amount shared as player A, subjects tend to estimate the shared amount rather accurately in all treatments.[30] In terms of the amounts shared back, participants tend to underestimate them substantially, especially in treatments without communication.[31] Intuitively, subjects believe that without communication their opponents share back very little and these amounts improve a lot with communication. In reality, individuals share back more generously without communication, while communication has either no or even a negative effect on trustworthiness.

[28] Two-sided MWU tests, $p = 0.067$ for *Base/NComm* vs. *NM/NComm*, $p = 0.183$ for *Base/NComm* vs. *M/NComm*, and $p = 0.779$ for *M/NComm* vs. *NM/NComm*.

[29] Two-sided MWU tests, $p = 0.054$ for *Base/NComm* vs. *NM/NComm*, $p = 0.538$ for *Base/NComm* vs. *M/NComm*, and $p = 0.194$ for *M/NComm* vs. *NM/NComm*.

[30] Using a two-sided matched pair Mann-Whitney U tests with continuity correction, we find no significant differences between the belief and the level of trust in *Base/NComm* ($p = 1$), *NM/NComm* ($p = 0.260$), *Base/Comm* ($p = 0.395$), *NM/Comm* ($p = 0.344$), *M/Comm* ($p = 0.359$) and only marginally significant differences in *M/NComm* ($p = 0.059$).

[31] Using a two-sided matched pair Mann-Whitney U tests with continuity correction, we find significant differences between the belief and the amount shared back in *Base/NComm* ($p = 0.006$), *M/NComm* ($p = 0.044$), *Base/Comm* ($p < 0.001$), *NM/Comm* ($p < 0,001$), *M/Comm* ($p < 0.001$) and marginally significant differences in *NM/NComm* ($p = 0.071$).

4.5 Additional insights

Differences between pro-choice and pro-life subjects. Overall, 146 subjects (48%) and 158 subjects (52%) indicated that they identify as pro-life and pro-choice, respectively. The remaining two subjects left the experiment before the questionnaire ended. Both subjects participated in *Base/Comm*.

Across all treatments, participants who identify as pro-life shared, on average, 5.32 out of 10 as Trustors. Participants who identify as pro-choice shared 5.96 out of 10 in the role of Trustors, and the difference between the two groups is not statistically significant (two-sided MWU test, $p = 0.112$). The average levels of trust are not different within treatments either,[32] with the exceptions of treatments *Base/Comm* and *NM/NComm* where participants who identify as pro-choice display higher levels of trust (two-sided MWU tests, $p = 0.040$ and $p = 0.098$ for *Base/Comm* and *NM/NComm*, respectively).

In the role of Trustees, pro-life and pro-choice subjects displayed average levels of trustworthiness of 6.97 and 6.80, respectively. The difference between the two groups is not statistically significant (two-sided MWU test, $p = 0.527$). There are no systematic differences between participants who identify as pro-life and pro-choice in terms of trustworthiness within treatments.[33]

Intensity of group membership preference and communication clusters. An *out-group* communication cluster yields detrimental effects on trust and trustworthiness. We analyze whether subjects with higher intensity of group membership are more likely to focus on their differences in communication. Surprisingly, this is not the case. In *M/Comm*, subjects in all communication clusters report, on average, similar intensity of their pro-life/pro-choice views. In particular, the intensity is 7.88 in *out-group*, 7.90 in *task*, 7.77 in *small talk*, and 8.21 in *no interaction*. The differences are not statistically significant[34] indicating that participants with higher intensity of group membership preference are not more likely to self-select into an *out-group* communication cluster.[35]

[32] The average levels of trust between participants who identify as pro-life and pro-choice are not statistically significantly different in *Base/NComm* ($p = 0.651$), *NM/Comm* ($p = 0.471$), *M/NComm* ($p = 0.739$) and *M/Comm* ($p = 0.221$). All p-values refer to the two-sided MWU tests.

[33] The average levels of trustworthiness between participants who identify as pro-life and pro-choice are not statistically significantly different in *Base/NComm* ($p = 0.948$), *Base/Comm* ($p = 0.685$), *NM/NComm* ($p = 0.877$), *NM/Comm* ($p = 0.769$), *M/NComm* ($p = 0.782$) and *M/Comm* ($p = 0.389$). All p-values refer to the two-sided MWU tests.

[34] One-sided MWU tests indicate that the intensity of group membership preference is not higher in the *out-group* communication cluster compared to *task* cluster ($p = 0.451$), *small talk* cluster ($p = 0.323$) and *no interaction* cluster ($p = 0.798$).

[35] The result looks similar once we focus on *NM/Comm*. Subjects in both communication clusters report, on average, similar intensity of their group identity. The intensity is 2.90 in *task* and 3.13 in *small talk*. The differences are not statistically significant ($p = 0.565$) using a one-sided MWU test. Thus, the data suggests that participants with higher intensity of group membership preference are not more likely to self-select into a less task oriented communication cluster.

5 Conclusion

This paper challenged a well-established finding that communication is universally beneficial in social dilemmas. We formalized a theoretical framework of a Trust Game featuring other-regarding preferences influenced by communication and social distance and experimentally put it to the test. In an online experiment, we captured the social distance between individuals via different group assignments. Using the treatments with no groups as a baseline (treatments *Base*), we first introduced the groups C and L without further meaning (treatments *NM*). Then, we added context to the groups, with groups C and L corresponding to pro-choice and pro-life individuals, respectively (treatments *M*). Furthermore, we varied whether individuals could communicate with each other (treatments *Comm* and *NComm*) for each type of group assignment. Our results address three pre-registered hypotheses along with additional exploratory findings.

The key contribution of our paper is providing a unique setup in which communication not only does not promote trust and trustworthiness but even harms these outcomes. When individuals are polarized enough, i.e. when one of them publicly identifies as pro-choice and another one as pro-life, communication no longer improves trust and even harms trustworthiness on average. Additionally, a large fraction of individuals focus their communication on their polarizing out-group characteristics leading to a substantial drop in levels of trust and trustworthiness in comparison to those who do not communicate at all.

Overall, the effects of communication on trust and trustworthiness differ drastically depending on the communication content. Using unsupervised machine learning, we established our participants follow one of four patterns in their communication: many focus on the task and discuss the best strategies, many engage in small talk about weather and a football game from last night, and some barely interact or do not interact at all. Moreover, when groups are polarized, many participants focus on their out-group differences in their communication. Importantly, this communication cluster only appears in treatment *M/Comm*. Focusing on the task is the most beneficial both in terms of trust and trustworthiness: it yields the most consistently positive boost across treatments for both outcomes. Small talk is helpful to improve trust unless participants are polarized enough (in treatment *M/Comm* but does not have a significant effect on trustworthiness. Focusing on the out-group characteristic results in a large-magnitude highly significant drop in both trust and trustworthiness in comparison to no communication.

We find mixed evidence for morality-based out-group discrimination. Without communication, morality-based discrimination does not occur to a significant extent for both trust and trustworthiness. However, with communication, both trust and trustworthiness are significantly harmed by polarization induced by morality-based groups in *M* compared to groups with super-

ficial meaning in *NM*.

Our paper has two key policy-relevant take-away messages: one is good and one is bad news. Let us start with the bad one. Interventions that feature direct communication should be treated with caution and assessed critically. As our results contain a "proof by counterexample" that the positive effect of communication is not bulletproof, it is important to take into consideration that polarization might not be the only aspect that shifts the communication impact to become negative. We speculate that other scenarios might include other types of out-group discrimination, interactions containing direct and indirect hostility, and so on.

The good news is, that even when individuals are polarized, many of them choose not to focus on their differences and work towards a joint goal instead. In this case, communication improves trust and trustworthiness, although not as strongly as without polarization. Of course, polarization reduction policies would be the most beneficial. However, even if softening polarization is not feasible, policies aiming at redirecting communication from out-group differences to focusing on the task could improve the outcomes substantially.

References

Aksoy, Billur, Haley Harwell, Ada Kovaliukaite, and Catherine Eckel (2018). "Measuring trust: A reinvestigation". *Southern Economic Journal* 84.4, 992–1000.

Aksoy, Ozan (2019). "Crosscutting circles in a social dilemma: effects of social identity and inequality on cooperation". *Social Science Research* 82, 148–163.

Andreoni, James and Justin M Rao (2011). "The power of asking: How communication affects selfishness, empathy, and altruism". *Journal of Public Economics* 95.7-8, 513–520.

Andres, Maximilian (2023). "Communication in the Infinitely Repeated Prisoner's Dilemma: Theory and Experimental Evidence". *arXiv preprint* 2304.12297.

Andres, Maximilian and Lisa Bruttel (2024). "Communicating Cartel Intentions". *Working Paper*.

Andres, Maximilian, Lisa Bruttel, and Jana Friedrichsen (2023). "How communication makes the difference between a cartel and tacit collusion: A machine learning approach". *European Economic Review* 152, 104331.

Babin, J Jobu and Haritima S Chauhan (2023). "Initiating free-flow communication in trust games". *Frontiers in Behavioral Economics* 2, 1120448.

Balliet, Daniel, Joshua M Tybur, Junhui Wu, Christian Antonellis, and Paul AM Van Lange (2018). "Political ideology, trust, and cooperation: In-group favoritism among Republicans and Democrats during a US national election". *Journal of Conflict Resolution* 62.4, 797–818.

Ben-Ner, Avner, Louis Putterman, and Ting Ren (2011). "Lavish returns on cheap talk: Two-way communication in trust games". *The Journal of Socio-Economics* 40.1, 1–13.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, reciprocity, and social history". *Games and Economic Behavior* 10.1, 122–142.

Bolton, Gary E and Axel Ockenfels (2000). "ERC: A theory of equity, reciprocity, and competition". *American Economic Review* 91.1, 166–193.

Burks, Stephen V, Jeffrey P Carpenter, and Eric Verhoogen (2003). "Playing both roles in the trust game". *Journal of Economic Behavior & Organization* 51.2, 195–216.

Charness, Gary, Giovanni Di Bartolomeo, and Stefano Papa (2024). "A stranger in a strange land: Promises and identity". *Games and Economic Behavior* 144, 13–28.

Charness, Gary and Martin Dufwenberg (2006). "Promises and partnership". *Econometrica* 74.6, 1579–1601.

Charness, Gary and Matthew Rabin (2002). "Understanding social preferences with simple tests". *The Quarterly Journal of Economics* 117.3, 817–869.

Danz, David, Lise Vesterlund, and Alistair J Wilson (2022). "Belief elicitation and behavioral incentive compatibility". *American Economic Review* 112.9, 2851–2883.

Dawes, Robyn M, Jeanne McTavish, and Harriet Shaklee (1977). "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation." *Journal of Personality and Social Psychology* 35.1, 1.

Ederer, Florian and Frédéric Schneider (2022). "Trust and promises over time". *American Economic Journal: Microeconomics* 14.3, 304–320.

Enke, Benjamin and Thomas Graeber (2023). "Cognitive uncertainty". *The Quarterly Journal of Economics* 138.4, 2021–2067.

Ewers, Mara and Florian Zimmermann (2015). "Image and misreporting". *Journal of the European Economic Association* 13.2, 363–380.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde (2023). "The preference survey module: A validated instrument for measuring risk, time, and social preferences". *Management Science* 69.4, 1935–1950.

Fehr, Ernst and Lorenz Goette (2007). "Do workers work more if wages are high? Evidence from a randomized field experiment". *American Economic Review* 97.1, 298–317.

Fehr, Ernst and Klaus M Schmidt (1999). "A theory of fairness, competition, and cooperation". *The Quarterly Journal of Economics* 114.3, 817–868.

Fiedler, Marina and Ernan Haruvy (2009). "The lab versus the virtual lab and virtual field—An experimental investigation of trust games with communication". *Journal of Economic Behavior & Organization* 72.2, 716–724.

Fiedler, Marina, Ernan Haruvy, and Sherry Xin Li (2011). "Social distance in a virtual world experiment". *Games and Economic Behavior* 72.2, 400–426.

Gächter, Simon, Eric J Johnson, and Andreas Herrmann (2022). "Individual-level loss aversion in riskless and risky choices". *Theory and Decision* 92.3, 599–624.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". *Journal of Economic Literature* 57.3, 535–574.

Glaeser, Edward L, David I Laibson, Jose A Scheinkman, and Christine L Soutter (2000). "Measuring trust". *The Quarterly Journal of Economics* 115.3, 811–846.

Goeree, Jacob K and Jingjing Zhang (2014). "Communication & competition". *Experimental Economics* 17.3, 421–438.

Grigoryan, Lusine, San Seo, Dora Simunovic, and Wilhelm Hofmann (2023). "Helping the ingroup versus harming the outgroup: Evidence from morality-based groups". *Journal of Experimental Social Psychology* 105, 104436.

Hossain, Tanjim and Ryo Okui (2013). "The binarized scoring rule". *Review of Economic Studies* 80.3, 984–1001.

Ismayilov, Huseyn and Jan Potters (2016). "Why do promises affect trustworthiness, or do they?" *Experimental Economics* 19, 382–393.

McCabe, Kevin A, Stephen J Rassenti, and Vernon L Smith (1998). "Reciprocity, trust, and payoff privacy in extensive form bargaining". *Games and Economic Behavior* 24.1-2, 10–24.

McCabe, Kevin A, Mary L Rigdon, and Vernon L Smith (2007). "Sustaining cooperation in trust games". *Economic Journal* 117.522, 991–1007.

Parker, Michael T and Ronnie Janoff-Bulman (2013). "Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love"". *Social Justice Research* 26.1, 81–96.

Petrishcheva, Vasilisa, Gerhard Riener, and Hannah Schildberg-Hörisch (2023). "Loss aversion in social image concerns". *Experimental Economics* 26.3, 622–645.

Rousseeuw, Peter J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20, 953–65.

Sheremeta, Roman M and Jingjing Zhang (2014). "Three-player trust game with insider communication". *Economic Inquiry* 52.2, 576–591.

Shu, Tao, Johan Sulaeman, and Eric P. Yeung (2012). "Local religious beliefs and mutual fund risk-taking behaviors". *Management Science* 58.10, 1779–1796.

Weisel, Ori and Robert Böhm (2015). ""Ingroup love" and "outgroup hate" in intergroup conflict between natural groups". *Journal of Experimental Social Psychology* 60, 110–120.

Xiao, Erte and Daniel Houser (2005). "Emotion expression in human punishment behavior". *Proceedings of the National Academy of Sciences* 102.20, 7398–7401.

A Proofs

A.1 Proof of player B's best response R^*

Proof. Let us derive the utility function of player B with respect to R.

$$\frac{\partial U_B(\pi_A(S,R),\pi_B(S,R),\beta)}{\partial R} = \frac{\partial}{\partial R}\left(\pi_B(S,R) - \beta \cdot \left(\frac{\pi_B(S,R)-\pi_A(S,R)}{\pi_A(S,R)+\pi_B(S,R)}\right)^2\right) \tag{9}$$

Rewrite both $\pi_A(S,R)$ and $\pi_B(S,R)$, and simplify.

$$\Leftrightarrow \frac{\partial U_B(\pi_A(S,R),\pi_B(S,R),\beta)}{\partial R} = \frac{\partial}{\partial R}\left(10 + 3\cdot S - R - \beta \cdot \left(\frac{4\cdot S - 2\cdot R}{20+2\cdot S}\right)^2\right) \tag{10}$$

Derive and set equal to 0.

$$\Leftrightarrow -1 + \beta \cdot 2 \cdot \left(\frac{4\cdot S - 2\cdot R}{20+2\cdot S}\right)\cdot \frac{2}{20+2\cdot S} = 0 \tag{11}$$

Add 1 and simplify.

$$\Leftrightarrow \left(\frac{4\cdot\beta}{(20+2\cdot S)^2}\right)\cdot(4\cdot S - 2\cdot R) = 1 \tag{12}$$

Multiply $\frac{4\cdot\beta}{(20+2\cdot S)^2}$ and simplify.

$$\Leftrightarrow 4\cdot S - 2\cdot R = \frac{(10+S)^2}{\beta} \tag{13}$$

Add $2\cdot R$, subtract $\frac{(10+S)^2}{\beta}$ and divide by 2.

$$\Leftrightarrow R = 2\cdot S - \frac{(10+S)^2}{2\cdot\beta} \tag{14}$$

Thus, the best response of player B is

$$R^*(S,\beta) = 2\cdot S - \frac{(10+S)^2}{2\cdot\beta} \tag{15}$$

This proofs Equation (2). □

A.2 Proof of player A's best response S^*

Proof. Let us derive the utility function of player A with respect to S given player B's best response R^*.

$$\frac{\partial U_A(S, R = R^*(S, \beta))}{\partial S} = \frac{\partial}{\partial S}\left(\pi_A(S, R = R^*(S, \beta))\right) \tag{16}$$

Rewrite $\pi_A(S, R = R^*(S, \beta)$ and simplify.

$$\Leftrightarrow \frac{\partial U_A(S, R = R^*(S, \beta))}{\partial S} = \frac{\partial}{\partial S}\left(10 + S - \frac{(10 + S)^2}{2 \cdot \beta}\right) \tag{17}$$

Derive and set equal to 0.

$$\Leftrightarrow 1 - \frac{2 \cdot (10 + S)}{2 \cdot \beta} = 0 \tag{18}$$

Add $\frac{2 \cdot (10+S)}{2 \cdot \beta}$ and multiply $2 \cdot \beta$.

$$2 \cdot \beta = 2 \cdot (10 + S) \tag{19}$$

Divide by 2 and subtract β.

$$S = \beta - 10 \tag{20}$$

Thus, the best response of player A given player B chooses her best response R^* is

$$S^*(\beta) = \beta - 10 \tag{21}$$

This proofs Equation (4). □

A.3 Proof of player B's best response $R^*(S, \beta)$ given player A best respond S^*

Proof. Let us plug in S^* into the best response of player B $R^*(S = S^*(\beta), \beta)$ to get the response of player B given player A best respond.

$$R^*(S = S^*(\beta), \beta) = 2 \cdot S^* - \frac{(10 + S)^2}{2 \cdot \beta} \tag{22}$$

Rewrite S^*.

$$\Leftrightarrow 2 \cdot \beta - 20 - \frac{(10 + \beta - 10)^2}{2 \cdot \beta} \tag{23}$$

Simplify.

$$\Leftrightarrow \frac{3}{2} \cdot \beta - 20 \tag{24}$$

Thus, the response of player B given player A's best respond is

$$R^*(S = S^*(\beta), \beta) = \frac{3}{2} \cdot \beta - 20 \tag{25}$$

This proofs Equation (5). □

A.4 Proof of interior solutions for $\beta > \frac{40}{3}$

Proof. Let us first calculate the β for which $R^*(S = S^*(\beta), \beta)$ yields an interior solution.

$$R^*(S = S^*(\beta), \beta) = \frac{3}{2} \cdot \beta - 20 > 0 \tag{26}$$

Solve for β.

$$\Leftrightarrow \beta > \frac{40}{3} \tag{27}$$

Thus,

$$R^*(S = S^*(\beta), \beta) = \frac{3}{2} \cdot \beta - 20 > 0 \quad \text{for} \quad \beta > \frac{40}{3} \tag{28}$$

This part of the proof establishes for $\beta > \frac{40}{3}$ where is an interior solutions to $R^*(S = S^*(\beta), \beta)$. Now, let us calculate the β for which $S^*(\beta)$ yields an interior solution.

$$S^*(\beta) = \beta - 10 > 0 \tag{29}$$

Solve for β.

$$\Leftrightarrow \beta > 10 \tag{30}$$

Thus,

$$S^*(\beta) = \beta - 10 > 0 \quad \text{for} \quad \beta > 10 \tag{31}$$

This part of the proof establishes for $\beta > 10$ where is an interior solutions to S^*. Taken together, as long as $\beta > \frac{40}{3}$, we obtain an inner solution for both $R^*(S = S^*(\beta), \beta)$ and $S^*(\beta)$.

□

A.5 Proof of Lemma 1

Let us derive the best response of player A $S^*(\beta)$ with respect to β.

$$\frac{\partial S^*(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}(\beta - 10) \tag{32}$$

Derive.

$$\Leftrightarrow 1 \tag{33}$$

Thus, the best response of player A increases in β:

$$\frac{\partial S^*(\beta)}{\partial \beta} > 0 \tag{34}$$

This proofs the first part of Lemma 1.

Let us derive the best response of player B $R^*(S = S^*(\beta), \beta)$ given player A best respond $S^*(\beta)$ with respect to β.

$$\frac{\partial R^*(S = S^*(\beta), \beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\frac{3}{2} \cdot \beta - 20\right) \tag{35}$$

Derive.

$$\Leftrightarrow \frac{3}{2} \tag{36}$$

Thus, the best response of player B $R^*(S = S^*(\beta), \beta)$ given player A best respond $S^*(\beta)$ increases in β:

$$\frac{\partial R^*(S = S^*(\beta), \beta)}{\partial \beta} > 0 \tag{37}$$

This proofs the second part of Lemma 1.

B Communication analysis

This section provides all the technical details for the communication content analysis.

B.1 Corpus

As input for our analysis, we use the chats from all treatments as our "corpus", considering one chat as one observation, i.e., "document". This corpus is subject to a systematic natural language processing procedure described in Section 4. Now, we transform this corpus into a term-frequency-inverse-document-frequency matrix $tfidf_{\alpha,\beta} = tf_{\alpha,\beta} \cdot idf_{\alpha}$. This matrix is a useful representation of a corpus to study the documents (see Gentzkow et al., 2019, and the literature therein). In this matrix, the rows denote documents and the columns denote unique tokens in this corpus. The entries in this matrix are calculated as the product of the term frequency $tf_{\alpha,\beta}$ and the inverse document frequency idf_{α}, where $idf_{\alpha} = \log_2 \frac{|D|}{|\{d|t_{\alpha} \in d\}|}$. D denotes the total number of documents in our corpus. $|\{d \mid t_{\alpha} \in d\}|$ denotes the number of documents in which the token t_{α} occurs. $tf_{\alpha,\beta}$ denotes how often the token t_{α} occurs in document d_{β}.

B.2 Hierarchical Cluster Analysis

After having computed the term-frequency-inverse-document-frequency matrix for our corpus, we cluster our corpus using a Hierarchical Cluster Analysis. We use an agglomerative approach to apply the unsupervised machine learning analysis. This approach works as follows: First, we treat each document as one cluster. Second, we repeatedly extend the clusters until all documents are in one cluster. At each iteration, the two less distant clusters form a new cluster. We use the minimum increase of the sum of squares method to determine which clusters determine a new cluster. The sum of squares between two clusters A and B is calculated as follows:

$$\frac{|A| \cdot |B|}{|A \cup B|} \|\mu_A - \mu_B\|^2, \tag{38}$$

where μ denotes the mean Euclidean distance.

B.3 Mean silhouette width

Figure A1 shows the mean silhouette width across the number of clusters k based on the dissimilarity matrix of the Hierarchical Cluster Analysis. The mean silhouette width measures how similar a document is to its own cluster compared to other clusters. The silhouette ranges from -1 to $+1$. A high value indicates that a document lies well within its own cluster (see Rousseeuw, 1987).

Figure A1 offers empirical support for our assumption that four is an appropriate number of clusters to group the communication content. The coefficients are reasonable and the elbow functional form suggests that our clustering structure matches the data well. This functional form suggests that four clusters work well. Thus, following this empirical support, we base the analysis on four clusters.

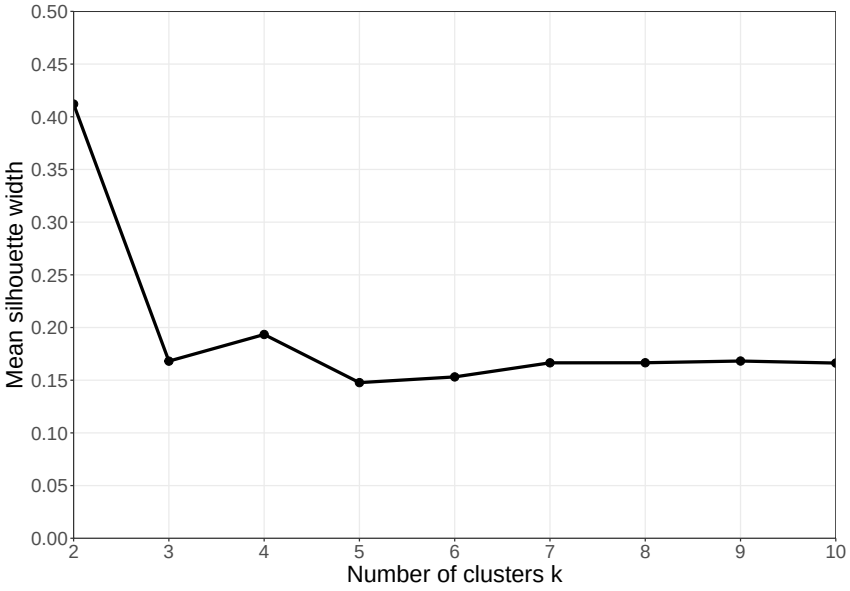


Figure A1: Mean silhouette width across number of clusters k based on the dissimilarity matrix.

C Instructions of the Experiment

C.0.1 Screen 1

Thank you for participating in this study.

This study will take approximately **15 minutes** to complete.

You will receive **a fixed payment of £2.25** for participating in this study, which will be paid upon completion independent of your decisions in the study. To complete the study, you will need to **read the following instructions carefully** and **answer the comprehension questions correctly**.

You have **a chance to earn an additional bonus of up to £21.24**. The amount of money you earn additionally strongly depends on your decisions. At the end of the study, **we pay the additional bonus to every tenth participant.**

To ensure you receive your participant payment, please **enter your Prolific ID** below. Your Prolific-ID:

C.0.2 Screen 2

Please read these instructions carefully. There will be comprehension checks.

This study consists of two parts: **Part A** and **Part B**. In this study, we calculate your earnings using points. At the end of this study, all your earnings will be converted from points to Pounds sterling using the following exchange rate:

$$\textbf{1 point} = \textbf{£0.36}.$$

After you complete Parts A and B, you will receive all feedback about your performance in Parts A and B and a detailed overview of the cumulative additional bonus you will have earned.

Please answer the following question. Please re-read the instructions above if you are not sure. You will have two opportunities to get this question correct.

How many parts does this study have?

- One part
- Two parts (correct answer)
- Three parts
- Four parts

C.0.3 Screen 3

Please read these instructions carefully. There will be comprehension checks.

Part A

In part A, you will interact with one other participant in the study.

You will interact in this task **once**. After you complete the task, we will ask you several questions about it. **Answering** these questions **truthfully may result in an additional monetary bonus.**

Please answer the following question. Please re-read the instructions if you are not sure. You will have two opportunities to get this question correct.
How many times are you going to interact with the other participant in this task?

- Once (correct answer)
- Twice
- Five times
- 50 times

C.0.4 Screen 4

Please read these instructions carefully. There will be comprehension checks.
Part A
Your task is as follows.

Each of you will be allocated 10 points.

Participant A can decide how many of these 10 points they want to share with Participant B. Participant B then receives **triple** the amount of points that Participant A shared.

Example. If Participant A shares 4 points with Participant B, it means that Participant A now has $10 - 4 = 6$ points and Participant B has $10 + 4 * 3 = 22$ points.

Participant B can then decide if they want to share their **total** amount of points (endowment and share from Participant A) with Participant A equally, or if they want to keep the total amount of points to themselves. Importantly, **sharing equally means** that Participant B splits the points such that **both Participants receive the same amount of points**.

Example. If Participant A shares 2 points with Participant B, it means that Participant A now has $10 - 2 = 8$ points and Participant B has $10 + 2 * 3 = 16$ points. If Participant B decides to

share all the points equally, it means their combined amount of points ($8 + 16 = 24$) are divided between both Participant A and Participant B equally. Therefore, Participant A gets $24/2 = 12$ points and Participant B gets $24/2 = 12$ points.

First, you will be asked to pick the amount you would share, given you are Participant A. Second, you decide for each amount you could have received from Participant A, if you would share the points equally or keep the points all to yourself, given you are Participant B. Finally, a random mechanism will decide, if you are Participant A or B. There is a **fifty-fifty chance that you will be Participant A or B.**

Please answer the following question. Please re-read the instructions above if you are not sure. You have two opportunities to get this question correct.
Consider you are Participant A and choose to share 4 points with Participant B. How many points will Participant B receive?

- 4 points
- 8 points
- 12 points (correct answer)
- 16 points

C.0.5 Screen 5

Please answer the following question. Please re-read the instructions below if you are not sure. You have two opportunities to get this question correct.
Consider that you are Participant B. Given that Participant A shared 2 points with you, how many points do you have when you choose to share all points equally?',

- 4 points
- 8 points
- 12 points (correct answer)
- 16 points

[The same instructions as on Screen 4 are here.]