The Influence of Social Interaction on Belief Biases

Andreas Grunewald[a] Victor Klockmann[b,c,d] Alicia von Schenk[b,c,d]
Ferdinand A. von Siemens[d,e,f,g]

February 27, 2024

Abstract

This paper examines the potential reinforcement of belief biases through social interaction when two individuals with similar biases communicate with each other. For this purpose, we propose a controlled laboratory experiment that allows for the manipulation of belief biases and the communication environment. Our findings indicate that communication, even among like-minded individuals, diminishes belief biases in the absence of external social cover for maintaining divergent beliefs. In the presence of social cover, however, communication does not reduce but rather exacerbate belief biases. Our empirical evidence suggests that social cover enables subjects to selectively disregard opinions that challenge their biases.

JEL: C91, C92, D83
Keywords: Belief bias, Social interaction, Motivated beliefs

---

[a]Frankfurt School of Finance and Management, Department of Economics, Adickesallee 32-34, 60322 Frankfurt, Germany.

[b]Julius-Maximilians-Universität Würzburg, Department of Economics, Sanderring 2, 97082 Würzburg, Germany.

[c]Center for Humans and Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.

[d]Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60323 Frankfurt, Germany.

[e]CESifo, Poschingerstrasse 5, 81679 Munich, Germany.

[f]Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany.

[g]Corresponding author. Email: vonsiemens@econ.uni-frankfurt.de. Phone: +49(69)79834824.

1 Introduction

A significant portion of our beliefs exhibit systematic distortions. We display overconfidence regarding our abilities or outward appearance (Moore and Healy; 2008; Kogan et al.; 2021; Huffman et al.; 2022), endorse fabricated statistics or false information about political outcomes, and hold self-serving beliefs about individuals outside our social circles (Di Tella et al.; 2015; Ging-Jehli et al.; 2020). Importantly, these (biased) beliefs are not formed in isolation but are often influenced and shaped by social interaction. This aspect has gained increasing attention because there is growing evidence that individuals tend to seek out communication partners who share similar beliefs and, arguably, similar belief biases (Bakshy et al.; 2015; Barnidge; 2017) – a tendency that appears to be reinforced by the matching algorithms of various social media platforms (Cinelli et al.; 2021). In light of this evidence, a major concern is that selective communication may aggravate belief biases rather than accumulate information, potentially undermining social cohesion by promoting extremism, polarization of political beliefs, violence, political gridlock, and social immobility (Levy and Razin; 2019; Sunstein; 2017).

Despite the severe consequences that a reinforcement of belief biases may have, causal evidence whether social interaction can indeed lead to an accumulation of belief distortions rather than information is still largely absent. This paper thus develops an experimental paradigm that facilitates the analysis how belief distortions are spread through social interactions. Embracing the notion that communication may be particularly problematic if belief distortions are shared within a communication network, we study two main questions: (i) Does communication among individuals with similar belief biases lead to a reinforcement of the biases? (ii) What kind of communication environments are particularly susceptible to cause a reinforcement of belief distortions?

Our experimental design has three important features which are essential to investigate these questions. First, the experiments take place online which allows us to implement social interaction between participants in a natural but controlled way through free-form chats. Second, the setup allows us to exogenously distort participants' beliefs without affecting the information they hold. This feature turns out to be indispensable for drawing inferences about whether communication among individuals with similar biases leads to a spread of the biases.

In particular, naturally occurring belief biases are typically correlated with preferences, information, communication habits, and how individuals process new information. Therefore, simply pairing participants who (naturally) hold similar belief biases would confound the effect of the belief bias with that of the individuals' characteristics and information. It, thus, would not be possible to distinguish whether social interactions lead to an accumulation of information or a reinforcement of biases. Third, we can exogenously change the communication environment to study what kind of environments are particularly susceptible to reinforce belief distortions.

In our experiment, we implement a simple decision environment in which participants are randomly assigned to groups of two, with one Player A and one Player B. Each group plays two dictator games, and each player is once the dictator and once the recipient. In the first dictator game, Player A is the dictator and can distribute an endowment either fairly or keep most of it for herself. In the second dictator game, Player B is the dictator and distributes another endowment. However, the exact options that Player B can choose from are randomly determined. In 50% of the cases, Player B has the same options as Player A – a fair split or keeping most for herself. In the other 50% of the cases, Player B has no options to choose from and the equal distribution is automatically implemented. This variation in choice options of Player B constitutes our first treatment dimension. The crucial element of the experiment is its information structure. All players are fully informed about the overall design, but they receive no information about the other players' choices. Further, Player A receives no information on which of the choice sets will materialize for Player B.

Our main object of interest is the belief that Player B holds about the behavior of the matched Player A. This belief should in principle not depend on Player B's own choice options, which are known to be unknown to Player A. The literature on motivated beliefs (Di Tella et al.; 2015; Zimmermann; 2020), however, suggests that it might be easier for Player B to take most of the endowment for herself if she beliefs Player A also did so. As generating a justification is only advantageous if the endowment can be split unevenly, Player B should only distort her beliefs in a self-serving manner when having this option available. In other words, the variation in choice options induces an exogenous distortion in the beliefs of Player B without providing information – a prerequisite for delineating the causal effect of social interactions on belief distortions.

Building upon this choice paradigm, we implement two additional randomly assigned treatment dimensions. First, we introduce free-form chats that allow some Players B to communicate with another Player B who faces the same choice options. In other words, participants that hold on average the same belief distortions communicate with each other. Second, we manipulate the communication environment. In social interactions, participants might be afraid to express or hold on to opinions that might be seen as unpopular (Loury; 1994; Morris; 2001; Braghieri; 2022; Golman; 2022). Based on the ideas of Masser and Phillips (2003) and Bursztyn, Egorov, Haaland, Rao and Roth (2023) the willingness to dissent crucially depends on the availability of social cover, i.e., other opinions, evidence or information that offer rationales for holding certain (maybe unpopular) beliefs. We vary social cover by displaying two opposing rationales to participants with respect to the behavior of Player A.

We start the analysis by focusing on the conditions without communication and without social cover. These conditions show that the variation in choice options for Player B indeed caused substantial belief distortions. In particular, the fraction of Players B believing that their Player A choose the unequal split increases from 42% if Players B do not face choice options to 62% if they do face choice options – an increase by roughly 50 percent. This 20pp difference in beliefs constitutes our measure for the prevalence of belief distortions of Players B who form their beliefs in isolation.

The second set of results addresses whether communication between individuals holding similar belief biases fosters the propagation of biases. We again first consider the conditions without social cover. In the conditions with communication, the fraction of Players B who have no choice option but believe Player A to have chosen the unequal split is 37%. This number increases to 44% for Players B that face a choice option and communicate with each other via free form chats. Hence, the belief distortions of Players B are significantly smaller with communication (7pp) than without communication (20pp). In line with literature that communication tends to improve choices in various contexts (**?**Cooper and Kagel; 2005; Kocher and Sutter; 2005), social interaction without social cover reduces belief biases in our setting.

Next, we repeat the same analysis in the conditions with social cover. In these conditions, thee effect of social interaction on belief biases reverses: communication leads to an increase in the measure of distortion in beliefs from 10pp without to 16pp with communication. While

the figures imply an economically relevant increase in distortions by 60%, the estimates are somewhat imprecise and therefore the difference between them turns out to be statistically insignificant. More importantly, our findings strongly emphasize the role of the communication environment in the formation and reinforcement of biased beliefs. In particular, the dissemination of belief biases via social interaction is significantly more pronounced under social cover than without cover.

Finally, to understand why communication under social cover is significantly more likely to reinforce belief biases, we analyze how communication differs between treatments with and without social cover. While our experiment is not designed to elucidate the exact mechanisms, the chat content and the correlation of beliefs between chat partners provide some suggestive evidence. Various analyses suggest that the content and tone of chats is not affected by social cover. Nevertheless, we find that the convergence of beliefs within chats is much stronger without social cover, suggesting that the influence of the chat partner on beliefs is stronger without cover. The chat content suggests that the weaker convergence of beliefs under social cover is particularly pronounced when the partner expresses an opinion that would contradict a potential belief bias. Our data thus indicate that social cover selectively disrupts the link between communication that corrects self-serving belief biases and the beliefs themselves.

2 Related Literature

There is a growing literature that discusses potential adverse effects of individuals' inclination to communicate with like-minded others. Levy and Razin (2019) and Sunstein (2017) argue that the endogenous selection of communication partners can not only result in political polarization but also contribute to political gridlock, hinder social mobility, and eventually pose a threat to democracy. These arguments inherently build on the idea that selected social interactions induce a proliferation of biases in beliefs and not just an accumulation of information. While there is amassing evidence that individuals actively select their communication network online (Bakshy et al.; 2015; Cinelli et al.; 2021) and offline (Barnidge; 2017), causal evidence on the extent to which these selected networks indeed reinforce belief biases instead of information is largely non-existent. Our paper aims to shed light on this issue. Complementing the studies above, we show that the effect of social interactions on belief biases is contingent upon the interaction environment. Specifically, social interactions are

significantly more likely to reinforce belief distortions under social cover than in the absence of social cover.

We also contribute to a recent discussion if and to what extent individuals react to information in the media more generally, and to cross-cutting news in particular. While Heatherly et al. (2017) conclude that cross-cutting information can help to moderate negative beliefs about opposing political parties, Bail et al. (2018) find evidence for the opposite effect. Connecting these two points of view, Levy (2021) shows that counter attitudinal news on Facebook decreases negative attitudes towards opposing political parties but does not affect political opinions. Our laboratory setup allows to distinguish how social interaction impact the proliferation of belief biases versus the proliferation of information. This distinction is important, because expected adverse consequences of selected communication networks are particularly severe if social interaction therein reinforces belief distortions and not just accumulates information.

Our findings also speak to an ongoing discourse surrounding the impact of fake news on social media platforms (Allcott and Gentzkow; 2017; Barrera et al.; 2020; Bursztyn, Rao, Roth and Yanagizawa-Drott; 2023). While previous studies primarily examine the direct effects of misinformation, our research uncovers an additional, indirect effect. Bursztyn, Egorov, Haaland, Rao and Roth (2023) argue that fake news can serve as social cover for opinions that may otherwise carry stigma. Embracing this notion, our paper shows that misinformation can contribute to an environment that fosters the reinforcement of biases through communication. These findings support arguments in favor of debunking fake news before they spread through social networks in order to avoid communication environments that are prone to reinforce the biases.

When investigating the impact of social interactions on belief distortions, our paper specifically focuses on distortions stemming from motivated reasoning. In doing so, we draw upon existing literature that has extensively documented instances of motivated reasoning in controlled laboratory settings (Di Tella et al.; 2015; Ging-Jehli et al.; 2020; Zimmermann; 2020; Drobner; 2022; Oprea and Yuksel; 2022), competitive debating environments (Schwardmann et al.; 2022), and among management professionals (Huffman et al.; 2022). Our choice to exploit motivated reasoning is twofold. Firstly, the body of work on motivated reasoning enables us to exogenously bias beliefs without altering individuals' information. Secondly,

motivated reasoning has been identified as a significant catalyst for political polarization and the formation of biased political opinions (Bénabou and Tirole; 2006; Bénabou; 2015; Levy and Razin; 2019). While our findings hold implications for the broader dissemination of belief biases, the latter argument underscores the importance of studying the dissemination of motivated beliefs as such.

Methodologically, a key contribution of our work is the exogenous manipulation of individuals' belief biases in a natural communication setting, that is, free-form chats. On the one hand, this design aspect allow us to cleanly measure the prevalence of bias, which is often challenging (Benoît and Dubra; 2011). On the other hand, it eliminates the confounding effects of naturally occurring biases as drivers of our results. For example, selective updating in social interactions by individuals with different levels of confidence, as in Oprea and Yuksel (2022), could be driven by differences in cognitive uncertainty (Enke and Graeber; 2019), social image concerns (Ariely et al.; 2009; Ewers and Zimmermann; 2015), communication habits, or private information. By attenuating these drivers of belief change, we can cleanly disentangle whether communication leads to the reinforcement of biases rather than the aggregation of information.

Last but not least, our paper contributes to an experimental literature that has studied how communication in groups affects choices in various contexts. Referring to some of the most prominent examples, papers have documented that communications makes group decisions display less risk aversion (Stoner; 1961; Teger and Pruitt; 1967), lead to more prosocial behavior (Cason and Mui; 1998; Bartling et al.; 2022), and make decisions reflect more closely the predictions of Nash Equilibrium (Bornstein and Yaniv; 1998; Cooper and Kagel; 2005; Kocher and Sutter; 2005). We complement these studies in two ways. First, our object of interest are the distortions in individuals' beliefs. Second, in our setup, all decisions are taken in complete isolation. Hence, we identify the pure effect of communication on the dissemination of belief biases, while excluding possibly confounding feelings of responsibility or social image concerns.

3 Experimental Design

Our study investigates in which environments communication between participants with biased beliefs propagates the bias. Studying this question requires a choice paradigm with

three essential features. First, we need to cleanly measure a meaningful, behaviorally relevant belief of participants in an incentivized manner. Second, the paradigm has to enable us to induce a bias in this belief exogenously. In particular, naturally occurring biases in beliefs inherently correlate with unobserved personal experiences, preferences, and economic circumstances. If communication occurred among participants that naturally hold the same belief biases, it would be impossible to delineate whether the communicating partner's biases, information, or personal characteristics drive posterior beliefs. Third, manipulating beliefs in the treatment group must not affect participants' information. If the treatment manipulation was changing information, it would not be possible to disentangle whether any induced shift in beliefs is due to a bias in beliefs or to the informational part of the treatment. This latter part is non-trivial because we most often associate belief changes with incoming information. Considering these necessary features for our experiment, the following paragraphs describe the implemented choice paradigm and treatments.

3.1 Choice Paradigm

We implement a simple decision environment in which participants are randomly matched into groups of two, with one Player A and one Player B. Each group plays two binary dictator games without feedback, and each player is once the dictator and once the recipient. In the first dictator game, Player A is the dictator and has two options to distribute an endowment of £5: she can either choose an equal split of £2.50 for each or allocate £4 to herself and £1 to Player B. In the second dictator game, Player B is the dictator with another, additional endowment of £5 to the two players. The exact options that Player B can choose from are randomly determined. In 50% of the cases, Player B faces the same options as Player A, that is, either an equal split of £2.50 for each or £4 to herself and £1 for Player A. In the remaining 50% of the cases, Player B has no option to choose from, and the allocation is automatically the equal split.

The crucial element of the experiment is its information structure. All players are completely and equally informed about the overall design of the experiment. However, they do not receive any information about the other players' choices or choice options. When making her choice, Player A knows neither the choice of Player B nor Player B's choice options. Equally, when making her choice, Player B does not know the choice of Player A. This information

structure is essential because it immediately implies that it is common knowledge that the choices of Player A cannot depend on the randomly determined and unknown choice options of Player B.

Our primary outcome measure in all treatments is the belief that Player B holds about the prior behavior of Player A. After her allocation choice, we ask Player B whether the specific Player A matched with her selected the equal split or not. If the answer is correct, Player B earns an additional £2.50. Our main outcome variable *"unfavorable belief"* indicates that Player B believes Player A to have chosen the unfair allocation.[1] Note that this belief is behaviorally meaningful because it will affect the decision of Player B, and we can cleanly measure it in an incentivized way – it thus fulfills the first requirement for the choice paradigm described above.

We implement the simple binary dictator games to create an obvious tension between self-gratification and social behavior for the participants. While our information structure implies that Player B's belief should not depend on her choice set, the literature on motivated beliefs (Di Tella et al.; 2015; Zimmermann; 2020; Drobner; 2022) suggests that it nevertheless does. In particular, it might be psychologically easier for Player B to take the £4 for herself if she believes Player A did the same. Generating such a justification to resolve the tension between self-gratification and social behavior self-servingly is only advantageous if the unfair option is available. Hence, Players B with no choice options and those with two choice options are identical on average, except that participants with the larger choice set have a stronger motivation to believe that Player A has been greedy. Any self-serving bias in beliefs induced by the treatment manipulation is therefore exogenous to subjects' characteristics and to their prior beliefs about the behavior of Player A.

The induction of motivated beliefs therefore fulfills requirements number two and three on the choice paradigm raised above – it is exogenous and uninformative.[2]

[1] We also elicited a more general belief. For this purpose, we ask Players B how many of the previous 100 Players A picked the unequal option. If the answer deviates by at most 5 in absolute terms from the correct answer, the subject earns £2.50. For this *"unfavorable general belief"*, we find similar but weaker and sometimes insignificant results (see Appendix B). This difference is consistent with Di Tella et al. (2015), who also find more substantial effects for specific than for general beliefs. It is also consistent with the idea of motivated beliefs: It is sufficient to hold an unfavorable belief about one's partner to justify unfair behavior.

Table 1: Treatment Overview

	No Cover		Social Cover	
	No Chat	Chat	No Chat	Chat
Unmotivated	UNMOT-NOCHAT	UNMOT-CHAT	UNMOT-NOCHAT-SOCO	UNMOT-CHAT-SOCO
	$n = 171$	$n = 324$	$n = 143$	$n = 270$
Motivated	MOT-NOCHAT	MOT-CHAT	MOT-NOCHAT-SOCO	MOT-CHAT-SOCO
	$n = 164$	$n = 333$	$n = 146$	$n = 270$

Note: The table provides an overview of the different treatments of the experiment and the number of Players B in each treatment after taking out the fastest 15% (see below). We balanced the number of independent observations and thus collected about twice as many subjects in the communication treatments as compared to the no communication treatments.

3.2 Treatments

Table 1 depicts an overview of the treatments implemented in our 2x2x2 design. As described above, our first treatment dimension varies the incentives for participants to hold motivated beliefs. Player B in the treatment arms labeled MOTIVATED has the option to allocate the endowment unevenly. Therefore, this participant is motivated to distort her beliefs regarding Player A's behavior. In contrast, Player B in the treatment arms labeled UNMOTIVATED has no choice options and, therefore, no incentive to distort her beliefs. From the perspective of Player A, these two treatments are identical because this player is not informed about the choice options of Player B. The average differences in beliefs between the MOTIVATED and UNMOTIVATED treatment arms measure the extent to which individuals in the MOTIVATED treatments hold biased beliefs.

Believing that the full population of participants is selfish is, however, not necessary.

[2]We also elicited the analogous beliefs of Player A as well as her second-order beliefs about Player B's beliefs. If Player B has no choice, we ask Player A only to report her second-order belief concerning Player B. In this case, we inform Player A that Player B has no choice after making her allocation choice.

Our second treatment dimension varies whether or not participants interact with another subject who holds the same motivation to bias her belief about Player A. Players B in the treatment arms labeled NoChat do not interact with any other participant during the experiment. Players B in the treatment arms labeled Chat enter a surprise communication stage immediately after reading the instructions and before making her allocation decisions (if any) and reporting their beliefs.[3] The free-form chat lasts for three minutes, and we encourage participants to discuss the previous behavior of Players A and their intended own decisions before the chat starts. At the communication stage, we group Players B together with the same motivation to distort their beliefs about their Players A; they are both either in a Motivated or an Unmotivated treatment arm. Therefore, participants in the conditions Unmot-Chat and Mot-Chat both communicate and are, on average, identical in all aspects except for their motivation to distort their beliefs. The difference in average beliefs between Unmot-Chat and Mot-Chat thus allows us to measure the extent to which individuals in Mot-Chat hold biased beliefs after communicating. In the same way, the difference between Mot-NoChat and Unmot-NoChat quantifies the biased beliefs of individuals without communication. Importantly, when we compare these two differences, we can identify to what extent communication with individuals holding similar biases propagates these biases. Therefore, when describing the results in Section 4, this difference-in-differences is a major outcome of interest.

Finally, we implement a third treatment dimension to analyze which characteristics of the communication environment foster or mitigate the propagation of belief biases. An essential aspect in this regard is whether or not participants are willing to share and uphold opinions they believe to be socially unpopular (Loury; 1994; Morris; 2001; Braghieri; 2022; Golman; 2022). Individuals appear to be more inclined to communicate unpopular beliefs if there is social cover, such as a rationale for a certain opinion (Bursztyn, Egorov, Haaland, Rao and Roth; 2023) or evidence that others hold a similar belief (Masser and Phillips; 2003). To create a communication environment which facilitates dissent and the expression of unpopular opinions, we thus provide such social cover in the treatment arms labelled SocialCover. In particular, we display two opposing quotes to all Players B after the instructions and before a potential chat stage. These quotes are

[3] A screenshot of the chat window can be found in Figure 3 in the appendix.

"I think we are living in selfish times."

Javier Bardem, Hollywood actor and Oscar winner.

and

"I'm just thankful I'm surrounded by good people."

Jon Pardi, singer and songwriter.

All participants in the conditions with SocialCover receive both quotes, also participants that do not have the opportunity to chat. The two quotes generate a plurality of opinions that should allow participants to find social cover for holding either opinion in the chat. While these – contradictory – quotes arguably do not push participants' beliefs in a particular direction, they provide justifications to voice unpopular opinions and help to reduce pressure to agree with opinions of their chat partner. In the conditions with NoSocialCover, we simply do not provide the quotes. Because we think of the treatments without the quotes as our baseline, we omit the label NoSocialCover in their acronyms for expositional clarity.

3.3 Experimental Procedures

At the beginning of the experiment, participants received written on-screen instructions explaining the rules and details of the experiment. Afterward, they answered control questions ensuring a basic understanding of the experiment. We excluded the 22% of registered participants that did not answer all of these questions correctly. We informed the remaining participants whether they were Player A or B. The main part of the experiment started once we matched each Player B to a chat partner by arrival time. Due to the large active subject pool, the average waiting time was only a few seconds. To hold waiting times between the introduction and the main part constant across treatments, we formed pairs of Players B in all conditions, although only those in the Chat conditions interacted with each other. At the end of the experiments, all participants answered short surveys on demographics, their education, and social media usage.[4] For the participants in the conditions with SocialCover,

[4]In terms of social media usage, we elicited whether participants actively create content on social media, how often they use social media, and whether they share their political views on social media.

we additionally elicited social preferences by standard questionnaire items.[5]

For all treatments, we first implemented sessions with only the participants in the role of Player A. We informed them that we would later match them with a participant in the role of Player B and that they would receive their payment resulting from the choices of Player B in a second tranche. After collecting the decision data from Players A, we ran sessions with the participants in the role of Player B. We matched each participant in these sessions with one Player A from the first part of the experiment. After this second part of the experiment, we transferred the remaining payoff from Player B's decision to the corresponding Player A.

Our main object of interest is the belief of Player B with respect to the behavior of the Player A that is matched with her. To obtain approximately the same number of independent observations in each treatment, we oversampled the treatments in the CHAT conditions by a factor of two. A potential problem with online experiments is that participants might click through the experiment without paying attention to the instructions. To improve data quality, we drop the fastest 15% of participants in each treatment, overall 316 observations. Our results are qualitatively robust to including these observations.

We conducted the experiment online using Prolific and oTree (Chen et al.; 2016). Participants took, on average, about 8 minutes to complete the experiment and earned on average £6.00. In total, we had 4316 participants in the experiment. We received prior ethics approval from the joint ethics committee of Goethe University Frankfurt and Johannes Gutenberg University Mainz. We conducted the experiment in two waves. The first wave elicited observations for all treatments in the conditions with NOSOCIALCOVER, and the second wave for all treatments in the conditions with SOCIALCOVER. Table 1 provides an overview of the number of observations for each treatment, including the abbreviations for the treatments that we use in the text. We held all key aspects of the experiment constant across waves. In particular, the order and content of the experimental tasks and instructions were identical across all participants, treatments, and waves of the experiment. Moreover, when recruiting participants, we imposed identical constraints on the characteristics of the participant pool. Table 5 in the appendix summarizes the balance checks showing that these procedures resulted

[5]In analogy to Falk et al. (2018), we measure positive reciprocity and general trust by asking, on eleven-point Lickert scales, whether participants are willing to return a favor and whether they believe that people have only the best intentions. To measure altruism, we asked how many of unexpectedly received £1000 they would be willing to donate to a good cause.

in an overall well-balanced sample of participants across all treatment cells depicted in Table 1 with two exceptions. Participants in the first wave are less likely to be female and more likely to be older than the median age of 32 years than in the second wave. To account for these wave-specific differences in our empirical analysis, we always present both the raw treatment differences and specifications controlling for age and gender.

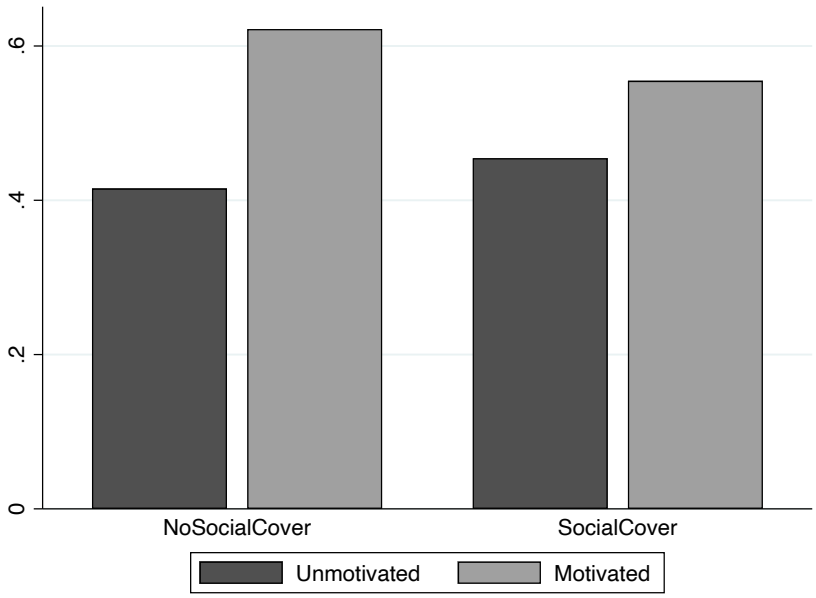
4 Results

In our empirical analysis, we first investigate biases in beliefs without social interaction by studying the difference between the MOTIVATED and UNMOTIVATED conditions in the NOCHAT treatment arms. We then analyze how communication with other participants holding a similar bias affects the prevalence of the bias. For this purpose, we study how the difference between the MOTIVATED and UNMOTIVATED conditions changes with social interaction, i.e., we employ a difference-in-differences approach. Finally, we explore what kind of communication environments are more or less susceptible to reinforce biases in beliefs, i.e., we investigate how the difference-in-differences depends on the availability of social cover in the communication environment.

4.1 Biases in Beliefs without Communication

A prerequisite for our analysis is that our experimental intervention exogenously shifts the beliefs of Player B concerning Player A. We test this presumption by comparing the MOTIVATED and UNMOTIVATED conditions in the scenarios without communication. Figure 1 shows the average beliefs of Player B in the four treatments without social interaction. It demonstrates that individuals in the MOTIVATED conditions hold strongly motivated beliefs. In particular, the share of Players B with unfavorable beliefs about their Players A increases from 42% in UNMOT-NOCHAT to 62% in MOT-NOCHAT. Similarly, the share of Players B with unfavorable beliefs about their Players A increases from 45% in UNMOT-NOCHAT-SOCO to 55% in UNMOT-CHAT-SOCO. Both differences in beliefs are statistically significant (Ranksum test, $p < 0.01$ and $p = 0.09$ in the NOSOCIALCOVER and SOCIALCOVER conditions, respectively). The effects are not only statistically significant but also economically relevant. In the treatments with NOSOCIALCOVER, the share of unfavorable beliefs increases

Figure 1: Average Unfavorable Beliefs Without Communication



Notes: The figure reports for the treatments without communication the fraction of Players B who believe that their Player A has chosen the unfair allocation. The dark gray bars report the fractions in the UNMOTIVATED and the light gray bars the fractions in the MOTIVATED conditions.

by roughly 20pp or 48% and in those with SOCIALCOVER by roughly 10pp or 22%. Overall, the treatment intervention MOTIVATED thus increases the share of participants attributing negative intentions to their partner: it induces a self-serving bias in participants' beliefs.

While the MOTIVATED conditions have a large effect on participants' beliefs in the absence of social interaction, this effect does not seem to differ across conditions with and without social cover. More specifically, Rank-sum tests show that beliefs do neither differ significantly between MOT-NOCHAT and MOT-NOCHAT-SOCO nor between UNMOT-NOCHAT and UNMOT-NOCHAT-SOCO (Ranksum-test, $p = 0.23$ for the MOTIVATED conditions and $p = 0.48$ for the UNMOTIVATED conditions). The contradictory quotes that we provide to participants thus do not appear to have a substantial effect on their beliefs about the behavior of Player A as such.

Result 1 *The beliefs of participants in the conditions with choice options are on average substantially biased in a self-serving fashion. This findings holds true for the conditions*

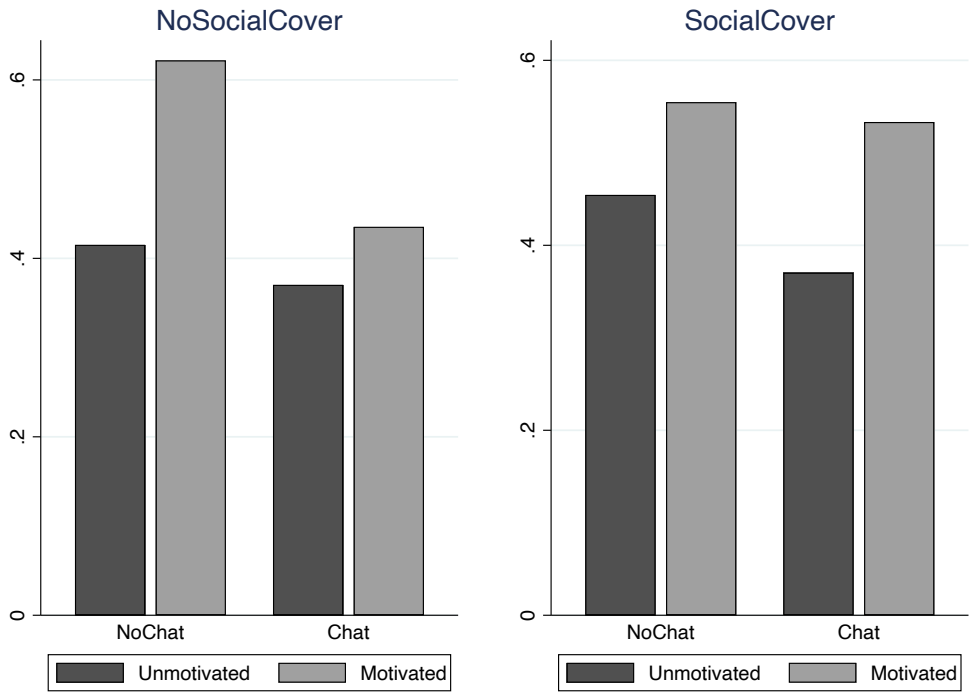
without and with social cover.

The identification of motivated beliefs builds on the idea of Di Tella et al. (2015) and Ging-Jehli et al. (2020), who argue that participants with the opportunity to take a larger share for themselves form more negative beliefs about their partner to justify their own choices. Our choice data are consistent with this argument. In both MOTIVATED conditions without social interaction, there is a strong correlation for Players B between holding an unfavorable belief about Player A's behavior and choosing an unfair allocation themselves. The respective Spearman correlations are 0.72 and 0.63 in MOT-NOCHAT and MOT-NOCHAT-SOCO ($p < 0.01$). In fact, across all conditions, around 90% of Players B that take the unfair action in MOTIVATED also hold the unfavorable belief that their partner previously took the unfair action.

4.2 Communication and Biases in Beliefs

As argued in the previous section, participants hold a substantial belief bias in the MOTIVATED conditions without social interaction. The main objective of our experiment is to investigate to what extent communication propagates this belief bias. Our results show that the effect of communication depends crucially on the environment in which social interaction takes place. Consider first the NOSOCIALCOVER conditions. The left panel of Figure 2 shows that communication without social cover substantially reduces the bias in beliefs. Participants still hold biased beliefs if they interact with others that hold a similar bias: the fraction of participants reporting an unfavorable belief about their partner increases from 37% in UNMOT-CHAT to 44% in MOT-CHAT. However, this increase by 7pp is less than half as large as the increase without communication. Communication without social cover – even among participants with similar belief distortions – thus attenuates rather than aggravates belief biases. Regression analysis presented in Column (1) of Table 2 confirms this finding. Taking individual unfavorable beliefs as the dependent variable, the estimate of the difference-in-difference coefficient is negative and statistically significant ($p = 0.04$). In line with the literature on choices in groups (**?**Cooper and Kagel; 2005; Kocher and Sutter; 2005), which documents that behavior in groups is more rational, we thus find that social interaction in the absence of social cover reduces biases in beliefs.

Figure 2: Average Unfavorable Beliefs



Notes: The figure reports for all treatments the fraction of Players B who believe that their Player A has chosen the unfair allocation. The dark gray bars report the fractions in the UNMOTIVATED and the light gray bars the fractions in the MOTIVATED conditions.

If we consider the conditions with SOCIALCOVER, the effect of social interaction on belief biases reverses. The right panel of Figure 2 illustrates that communication with social cover propagates the belief bias. The difference in average beliefs of Players B between MOT-NOCHAT-SOCO and UNMOT-NOCHAT-SOCO is 10pp, implying that participants hold biased beliefs in MOT-NOCHAT-SOCO. If participants communicate, the difference between the MOTIVATED and the UNMOTIVATED condition increases by 60% to 16pp. Accordingly, the difference-in-difference estimator in Column (2) of Table 2 is positive. While the increase of 60$ is substantial in economic terms, the regression results show that it turns out to be statistically insignificant ($p = 0.40$). In contrast to the results without social cover, communication thus does not attenuate biases if there is social cover. If anything, communication aggravates the biases in beliefs.

Columns (3) and (4) speak to our second main research question, i.e., which characteristics of

Table 2: Regression Results Unfavorable Beliefs

	NoSocialCover	SocialCover	All	
	(1)	(2)	(3)	(4)
Motivated	0.21***	0.10*	0.21***	0.21***
	(0.05)	(0.06)	(0.05)	(0.05)
Chat	-0.04	-0.08	-0.04	-0.04
	(0.05)	(0.05)	(0.05)	(0.05)
Chat × Motivated	-0.14**	0.06	-0.14**	-0.14**
	(0.07)	(0.07)	(0.07)	(0.07)
Social Cover			0.04	0.04
			(0.06)	(0.06)
Social Cover × Motivated			-0.11	-0.11
			(0.08)	(0.08)
Social Cover × Chat			-0.04	-0.04
			(0.07)	(0.07)
Social Cover × Chat × Mot			0.20**	0.20*
			(0.10)	(0.10)
Older Than 32 Years				-0.01
				(0.02)
Female				-0.08***
				(0.02)
Constant	0.42***	0.45***	0.42***	0.46***
	(0.04)	(0.04)	(0.04)	(0.04)
Number of Observations	992	829	1821	1821
adjusted R^2	0.03	0.02	0.02	0.03

Notes: The table reports the results of OLS regressions. The dependent variable is an indicator whether Player B holds the unfavorable belief on the behavior of Player A in all specifications. Column (1) analyses the NoSocialCover conditions, column (2) the SocialCover conditions, and columns (3) and (4) all data jointly. We report in parenthesis the standard errors clustered at the chat level for those who chat. Stars indicate significance at the 1%, 5%, and 10% level.

the communication environment tend to reinforce rather than mitigate belief biases. Indeed, we find that the effect of social interaction on the dissemination of biases differs significantly between the NoSocialCover and the SocialCover conditions. The coefficient of the triple interaction in Column (3) is positive and statistically significant ($p = 0.046$). Column (4) of Table 2 confirms that this result also holds if we control for the age and gender of participants. In fact, all coefficients literally remain identical after rounding to two digits.[6] Hence, communication environments that are characterized by a plurality of opinions that offer social cover for holding certain maybe unpopular beliefs are more prone to cause a reinforcement of belief biases than other environments in which social cover is absent.

Result 2 *In the conditions without social cover, communication among participants with similar belief biases reduces these biases. In the conditions with social cover, communication does not reduce but rather exacerbates belief biases. Consequently, the dissemination of belief biases by communication is significantly more pronounced with social cover than without social cover.*

4.3 Mechanisms

Result 2 documents that the characteristics of the communication environment are crucial for the impact of social interaction on belief biases. This result immediately raises the question how communication itself and participants' reaction to it differ between the environments with and without social cover. In general, there are two possible reasons why the effect of social interaction on belief biases may differ across the two environments. First, social cover may influence participants' attitudes toward communication and thus the content or tone of the chats. In particular, they might be less willing to communicate unfavorable beliefs about the behavior of Player A if these beliefs are biased against a social norm of prosocial behavior. With social cover, however, participants might be more likely to express biased unpopular beliefs, facilitating communication that reinforces belief biases. Second, social cover may change subjects' response to the chat content. Specifically, it may allow participants to hold

[6] In further regressions not reported here in detail, we fully interact the model with age or gender to test whether motivated beliefs and the effects of communication and social cover on the former might depend on these characteristics. None of these interactions are statistically significant ($p > 0.20$). While the statistical power of these regressions is lower, our coefficients of interest are again almost identical to the ones in Table 2.

biased, self-serving beliefs even when their chat partner holds and communicates a contrary point of view. In line with Oprea and Yuksel (2022), social cover may then facilitate selective updating such that participants adjust their beliefs more strongly when statements in the chat are consistent with their self-serving biases.

While our experiment is not designed to quantify the extent to which each of these reasons is driving our results, below we investigate the chat content to gather evidence for these potential mechanisms. Because we are interested in how social cover affects social interaction, we pool the data from Mot-Chat and Unmot-Chat and compare the chat content to the one in the conditions with social cover (Mot-Chat-SoCo and Unmot-Chat-SoCo). In analogy to above, we drop all observations from chats in which at least one of the chat partners belongs to the fastest 15% of participants in their treatment.

4.3.1 Willingness to Reveal Unfavorable Beliefs

We first study how social cover affects the content and tone of the chats. We compare the chat content and tone across conditions with four different approaches but we do not find any significant differences between the scenarios SocialCover and NoSocialCover. First, we hired two research assistants who coded the content in various aspects. Most importantly, they recorded whether a participant mentioned an unfavorable belief about the behavior of her Player A, and whether such an unfavorable belief was the first belief to be mentioned in the chat.[7] Note that these mentioned beliefs do not have to, and often do not, coincide with the incentivized stated beliefs after the chat. We find no significant difference in the number of times the unfavorable belief is mentioned or in the number of times the unfavorable belief is the first to be mentioned in a chat ($p > 0.40$, see Table 12). Second, we defined two word lists, which include words indicating that a subject expresses a favorable or unfavorable belief (see Appendix C.1 for details). For both lists, there are no significant differences in the number of chats that contain at least one word from the respective list across conditions with and without social cover ($p > 0.44$, Chi-squared tests). Third, we ran bigram and trigram

[7]Both research assistants first coded the chats independently, on the chat and individual levels. There was a high degree of agreement in the relevant variables between the two codings (the same coding in approximately 80% of all chats, Cramér's V above 0.7). Afterward, we asked them to provide a consolidated version of the coding by resolving any differences in their coding via discussion.

analyses, capturing the chats' most frequently mentioned pairs and triples of words. After merging similar word combinations, we ended up with three topics – splitting fairly, thinking about others, and wishing good luck (for more details see Appendix C.2). Again, there are no significant differences in the frequency with which either of the three topics is mentioned across the conditions Social Cover and No Cover ($p > 0.10$, Chi-squared tests). Fourth, there is no significant difference in the length of chats in terms of the number of words ($p = 0.17$, Kolmogorov–Smirnov test). Together, these analyses strongly suggest that social cover neither substantially changes the content of the chat nor the tone of communication.

Result 3 *The content or tone of the chats does not differ between the conditions with and without social cover.*

4.3.2 Immunity to Inconvenient Opinions

Next, we analyze how participants respond to the chat content. First, consider the correlation of stated beliefs within chat groups. If participants did not react to the observed chat, their beliefs should be uncorrelated within chat groups. Instead, we find that beliefs are correlated within chat groups in all four Chat treatments. In the NoSocialCover conditions, the Spearman correlation coefficients are 0.42 in Mot-Chat and 0.41 in Unmot-Chat ($p < 0.001$ for both correlations). In the SocialCover conditions, the Spearman correlation coefficients are 0.27 in Mot-Chat-Cover and 0.15 in Unmot-Chat-Cover ($p < 0.02$ for both correlations). Overall, the strong correlations between chat partners' beliefs shows that communication induces a convergence of beliefs among chat partners.

However, the correlation of chat partners' beliefs is weaker in the SocialCover conditions than in the NoSocialCover conditions. To confirm this observation, Table 3 shows linear regressions with the subject's stated beliefs as the dependent variable. Indeed, columns (3) and (4) show that social cover significantly reduces the correlation between chat partners' beliefs. The lower correlations imply that participants are more likely to stick to their own beliefs rather than adopting their chat partner's point of view when there is social cover – a finding consistent with Bursztyn, Egorov, Haaland, Rao and Roth (2023), who argue that social cover can lead to more dissent in the population.

Table 3: Regression Results Belief Correlations

	NoSocialCover	SocialCover	All	
	(1)	(2)	(3)	(4)
Other's Belief	0.43***	0.23***	0.43***	0.43***
	(0.05)	(0.06)	(0.06)	(0.05)
Social Cover			0.07	0.06
			(0.05)	(0.05)
Other's Belief × Social Cover			-0.20**	-0.21**
			(0.08)	(0.08)
Older than 32 Years				-0.04
				(0.04)
Female				-0.10**
				(0.04)
Constant	0.27***	0.34***	0.27***	0.35***
	(0.03)	(0.04)	(0.03)	(0.05)
Number of Observations	305	246	551	551
adjusted R^2	0.18	0.05	0.12	0.13

Notes: The table reports the results of OLS regressions. The dependent variable is an indicator whether Player B holds the unfavorable belief on the behavior of Player A in all specifications. Column (1) analyses the conditions without social cover, Column (2) the conditions with social cover, and Column (3) and (4) all data jointly. Other's Belief is the belief of the chat partner. We need not cluster the standard errors at the chat level because we take only one observation per chat. Stars indicate significance at the 1%, 5%, and 10% level.

Finally, we use our coding of chat content to further investigate participants response to expressed opinions in the social interaction. In particular, Table 4 correlates the belief that a participant states in the incentivized elicitation after the chat with whether or not her partner communicated an unfavorable belief in the chat. This correlation is significantly weaker in the SocialCover conditions than in the NoSocialCover conditions. Importantly, this reduction in the correlation seems to be entirely driven by cases in which the chat partner expresses an opinion that does not support the self-serving belief bias. Participants

Table 4: Stated Beliefs Conditional on Chat Partner's Mentioned Beliefs

	Partner mentioned unfavorable belief?	
	No	Yes
NoSocialCover	26.15%	67.16%
SocialCover	35.16%	64.61%

Notes: The table presents the likelihood of stating an unfavorable belief about the behavior of Player A conditional on social cover and on whether the chat partner mentioned an unfavorable belief in the chat.

in SocialCover seem to devalue their chat partners' opinions when these are inconvenient. Specifically, without social cover, only 26% of participants hold an unfavorable belief ex post when their chat partner mentions no or even a favorable belief (see Table 4). This number increases to 35% in the SocialCover treatments. As shown in Table 12 in the appendix, the reduction in correlation is statistically significant if the chat partner expresses an opinion that does not support the bias ($p = 0.038$) but there is no reduction in correlation otherwise. Overall, social cover thus seems to break the link between statements in the chat that correct self-serving biases and participants' ex-post beliefs.

Result 4 *Participants in the conditions with social cover are less responsive to communication, especially if the observed opinions inconveniently do not coincide with their self-serving belief bias.*

5 Conclusion

This paper presents a controlled online experiment studying the effects of social interaction on the formation of biased beliefs. Our findings reveal that communication without social cover reduces the bias in subjects' beliefs even if communication takes place among like-minded individuals. Communication with social cover, however, allows biases to persists or even reinforces biases. This finding highlights the important role of the communication environment on the proliferation of biases. Our evidence indicates that social cover enables individuals to selectively ignore information that does not support their desired beliefs.

These findings may have important implications for the regulation of news outlets and social media platforms. Our results suggest that the presence of a wide range of opinions may provide social cover that allows individuals to selectively ignore information that challenges their existing biases. Consequently, a plurality of opinions may contribute to the spread of biased beliefs. Social media platforms, and society in general, may therefore have a reason to regulate certain opinions, particularly extreme opinions based on fake news that are not helpful in forming factually correct public opinions. However, such interventions also threaten freedom of expression.

An important advantage of our setting is that it allows exogenous manipulation of social beliefs in a simple experimental paradigm. It is therefore well suited to study how biases in beliefs, rather than information, are spread through social networks. While our paper focuses on the spread of biased beliefs through communication in two-person chats, this is only one form of social interaction. Extending this analysis to multi-person chats, forums, endogenously selected communication partners, other forms of belief bias, or chat bots seems to be a rich, largely unexplored, and important area for future research. Such future research will hopefully provide guidance on how to strike a balance between countering the spread of belief bias and preserving the plurality of opinions and the vital freedom of expression that is essential for a healthy democratic discourse and a liberal society.

Appendix A Randomization Checks

We conducted the experiment in two waves: first the NoSocialCover conditions and then the SocialCover conditions. Every subject participated in only one role and one treatment. We ran all NoSocialCover conditions between the 30th of November and 16th of December 2021, and all SocialCover conditions between the 5th and 13th of December 2022. Table 5 shows the summary statistics of Players B in all treatments. To facilitate the exposition of the randomization checks and the ensuing statistical analysis, we binarize some variables that we did not elicit as binary variables. Concerning gender, 2% of our participants report to be neither male nor female, and we pool those with those who report being male. Median age in our sample is 33 years, and we create a dummy variable "Old" that indicates whether a participant is older than 32 years. 39% of our participants have at most a high school degree, which we indicate with our dummy variable "No University Degree". 50% of our participants report using social media daily, so we generate the corresponding dummy variable "Daily Social Media Use". The p-values in the last column refer to Chi-squared tests testing the null hypothesis of no differences across the eight treatments in both waves.

Table 5 shows that there are significant differences across treatments only in gender and age. These differences stem from the second wave containing the SocialCover treatments, in which participants are older and less likely to be female. To account for these wave-specific differences in our empirical analysis, we always present both the raw treatment differences and specifications controlling for age and gender.

Table 5: Summary Statistics

	NoSocialCover				SocialCover				
	Unmotivated		Motivated		Unmotivated		Motivated		
	NoChat	Chat	NoChat	Chat	NoChat	Chat	NoChat	Chat	p-value
Socio-Demographics									
Female	0.51	0.55	0.49	0.59	0.43	0.41	0.43	0.40	0.00
Older Than 32 Years	0.46	0.49	0.54	0.47	0.61	0.67	0.55	0.59	0.00
No University Degree	0.38	0.40	0.37	0.43	0.41	0.41	0.42	0.35	0.63
Social Media and Peers									
Daily Social Media Use	0.49	0.49	0.48	0.50	0.50	0.49	0.51	0.53	0.98
Creates Content on Social Media	0.25	0.30	0.27	0.28	0.24	0.34	0.30	0.29	0.41
Shares Political Views on Social Media	0.26	0.30	0.27	0.25	0.23	0.29	0.19	0.29	0.23
Similar Political Orientation as Friends	0.73	0.76	0.77	0.74	0.77	0.72	0.77	0.76	0.83
Number of Observations	171	324	164	333	143	270	146	270	

Notes: The table reports summary statistics for Players B in all treatments. The p-values refer to Chi-squared tests testing whether the distributions of variables are identical across treatments.

Appendix B Results on General Beliefs

In this section, we report the results concerning a more general unfavorable belief. In particular, we asked Players B how many of the previous 100 Players A picked the unequal option. If the answer deviated by at most 5 in absolute terms from the correct answer, the subject earned £2.50. For this belief, we find similar but overall weaker patterns as with the more specific beliefs. Considering the NoChat conditions, we find that the average general unfavorable beliefs increase from 59 in Unmot-NoChat to 63 in Motivated-NoChat and from 57 in Unmot-NoChat-SoCo to 62 in Mot-NoChat-SoCo. While these findings also indicate biased general beliefs, Rank-sum tests show that both increases are not statistically significant (p-value of 0.22 and 0.13). Considering the Chat conditions, we find that the average general unfavorable beliefs increase from 51 in Unmot-Chat to 55 in Mot-Chat and from 53 in Unmot-Chat-SoCo to 60 in Mot-Chat-SoCo. Communication thus has no effect on beliefs in the NoSocialCover conditions but slightly increases biases in the SocialCover conditions.

Regression analysis confirms these findings. Table 6 summarizes the results of OLS regressions with the general unfair beliefs, following the structure of the respective table in the main text. We find weak evidence for motivated beliefs in this variable as depicted in Column (1) and (2). The coefficients for motivated beliefs without communication are small and only significant in the NoSocialCover conditions (p-value of 0.10 and approximately 0.18). Social interaction slightly reduces biases in beliefs in the NoSocialCover conditions, and slightly increases biases in beliefs in the SocialCover conditions, but the interactions are not statistically significant (p-values larger than 0.66). The triple interaction measuring how social cover affects the effect of communication is positive but not significant (p-values larger than 0.71). Controlling for age and gender has very little effect on our coefficients of interest. Overall, the results for our general beliefs are therefore similar to but much weaker than the results for our specific beliefs. This difference is consistent with Di Tella et al. (2015), who also find more substantial effects for specific than for general beliefs. It is also consistent with the idea of motivated beliefs: It is sufficient to hold an unfavorable belief about one's partner to justify unfair behavior. Believing that the full population of participants is selfish is, however, not necessary.

Table 6: Regression Results General Unfavorable Beliefs

	NoSocialCover	SocialCover	All	
	(1)	(2)	(3)	(4)
Motivated	3.91	4.96*	3.91	3.88
	(2.93)	(2.97)	(2.93)	(2.94)
Chat	-7.39***	-3.31	-7.39***	-7.08**
	(2.75)	(2.86)	(2.75)	(2.75)
Chat × Motivated	-0.24	1.76	-0.24	-0.01
	(3.82)	(3.94)	(3.82)	(3.81)
Social Cover			-1.93	-2.20
			(3.03)	(3.03)
Social Cover × Motivated			1.05	1.00
			(4.17)	(4.17)
Social Cover × Chat			4.08	3.71
			(3.96)	(3.94)
Social Cover × Chat × Mot			2.00	1.69
			(5.49)	(5.46)
Older than 32 Years				-1.26
				(1.25)
Female				-6.12***
				(1.24)
Constant	58.64***	56.71***	58.64***	62.33***
	(2.10)	(2.18)	(2.10)	(2.30)
Number of Observations	992	829	1821	1821
adjusted R^2	0.02	0.01	0.02	0.03

Notes: The table reports the results of OLS regressions of treatment differences in the general unfair beliefs of Players A. Column (1) analyses the NoSocialCover conditions, column (2) the SocialCover conditions, and columns (3) and (4) all data jointly. We report in parenthesis the standard errors clustered at the chat level for those who chat. The stars indicates significance at the 1%, 5%, and 10% level, respectively.

Table 7: Distribution of Manual Topics in Chats

	Topic(s)			
Treatment	Fair	Unfair	Both	Number of Chats
NoSocialCover	265 [89.23%]	189 [63.64%]	169 [56.90%]	297
SocialCover	212 [86.89%]	164 [67.21%]	139 [56.97%]	244

Notes: The table reports the distribution of manually defined topics conditional on social cover. Relative frequencies per treatment reported in brackets.

Appendix C Code-Based Chat Analysis

Additional to the manual chat coding of the two research assistants, we conducted two code-based analyses. This section documents the details of both approaches in order to investigate how social cover impacts the content and tone of communication in our setting.

Appendix C.1 Results on Word Lists

In the first approach, we defined word lists to capture expressions indicating that a subject expresses a favourable or an unfavorable belief about the behavior of Player A. The word list to quantify positive expressions contains the following words: "fair", "fairly", "equal", "equally", "even", "evenly", "generous", "nice", "half", "kind", "split", "good", "hope". In analogy the word list indicating negative beliefs contains: "unfair", "unfairly", "greedy", "selfish", "keep", "kept", "take", "himself", "herself", "themselves", "bad". Table 7 reports the absolute and relative frequencies of both topics in the chats, conditional on whether or not social cover is available. There are no significant differences in these frequencies across the conditions Social Cover and No Cover ($p > 0.44$, Chi-squared tests).

Appendix C.2 Results on Bigram and Trigram Analysis

In order to quantify the tone and content of the chats, we also ran bigram and trigram analyses. These analyses capture the most frequently mentioned pairs and triples of words

Table 8: Bigram and Trigram Analyis of Chat Content – Topics

Bigram / Trigram	Count	Topic
50 50	142	Fair split
think participant	137	Thinking about others
participant chose	90	Thinking about others
think people	60	Thinking about others
good luck	56	Wishing good luck
split evenly	55	Fair split
even split	46	Fair split
think chose	46	Thinking about others
think participant chose	44	Thinking about others
chose 50	40	Fair split
think would	40	Thinking about others

Notes: The table shows the ten most frequent bigrams and trigrams that occur in the chats across all treatments. Column "Count" refers to the number of chats in which the bi-/trigram occurs. Column "Topic" refers to the topic assignment that was done manually after the identification of the bi-/trigrams.

in the chats (after deleting punctuation and stop words such as articles, prepositions, etc.). The top 11 (due to a tied 10th place) combinations are reported in Table 8 in descending order of frequency. After merging similar bi-/trigrams such as "50 50" and "even split" or "think participant" and "think people", we ended up with three topics – the fair split, thinking about others, and wishing good luck (see last column of Table 8). Table 9 reports the absolute and relative frequencies of the topics in chats, conditional on the availability of social cover. There are no significant differences in these frequencies across the conditions Social Cover and No Cover ($p > 0.10$, Chi-squared tests).

Table 9: Bigram and Trigram Analyis of Chat Content – Distribution of Topics

	Topic			
Treatment	Fair split	Thinking about others	Wishing good luck	Number of Chats
NoSocialCover	164 [55.22%]	234 [78.79%]	37 [12.46%]	297
SocialCover	119 [48.77%]	183 [75.00%]	19 [7.79%]	244

Notes: The table reports the distribution of topics from bigrams and trigrams conditional on social cover. Relative frequencies per treatment reported in brackets. Topics include bi-/trigrams as listed in Table 8.

Appendix D Additional Tables and Figures

Table 10: Regression Results Incentives and Belief Correlations

	NoSocialCover		SocialCover	
	(1)	(2)	(3)	(4)
Motivated	-0.02	-0.03	0.07	0.08
	(0.07)	(0.07)	(0.08)	(0.08)
Other's Belief	0.42***	0.41***	0.15	0.16*
	(0.08)	(0.08)	(0.09)	(0.09)
Other's Belief × Motivated	-0.00	0.01	0.12	0.11
	(0.11)	(0.11)	(0.13)	(0.13)
Older than 32 Years		0.02		-0.02
		(0.05)		(0.06)
Female		-0.09*		-0.12*
		(0.05)		(0.06)
Constant	0.24***	0.28***	0.34***	0.39***
	(0.05)	(0.06)	(0.06)	(0.07)
N	305	305	246	246
adjusted R^2	0.17	0.17	0.06	0.07

Notes: The table reports the results of OLS regressions. The dependent variable is an indicator whether Player B holds the unfavorable belief on the behavior of Player A in all specifications. Columns (1) and (2) analyses the conditions without social cover, columns (3) and (4) the conditions with social cover. Other's Belief is the belief of the char partner. We need not cluster the standard errors at the chat level because we take only one observation per chat. Stars indicate significance at the 1%, 5%, and 10% level.

Table 11: Differences in Mentioned Belief by Participant and First Mentioned Beliefs in Chat

	Participant mentions unfavorable belief?	First belief mentioned in chat was unfavorable?
	(1)	(2)
Social Cover	0.02	0.04
	(0.04)	(0.04)
Older than 32 Years	-0.03	-0.05*
	(0.03)	(0.03)
Female	-0.06*	-0.01
	(0.03)	(0.03)
Constant	0.39***	0.37***
	(0.03)	(0.04)
Number of Observations	1082	1082
adjusted R^2	0.00	0.00

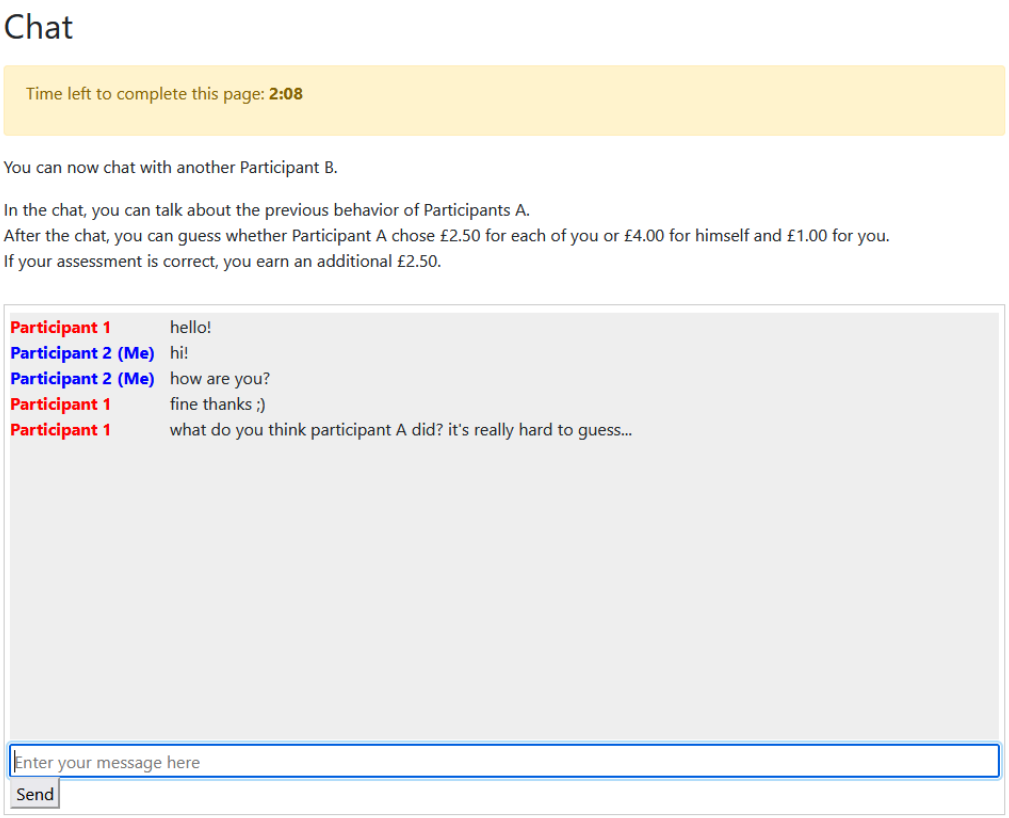
Notes: The table reports the results of OLS regressions. The dependent variable in Column (1) is whether the participant mentions the unfavorable belief in the chat. The dependent variable in Column (2) is whether the first mentioned belief in the chat was the unfavorable belief. We report in parenthesis the standard errors clustered at the chat level. Stars indicate significance at the 1%, 5%, and 10% level.

Table 12: Correlation of Stated Beliefs with Partner's Mentioning of Beliefs

	Partner mentioned unfavorable belief?	
	No	Yes
	(1)	(2)
Social Cover	0.07*	-0.03
	(0.04)	(0.05)
Older than 32 Years	-0.01	-0.10**
	(0.03)	(0.05)
Female	-0.11***	-0.06
	(0.04)	(0.05)
Constant	0.32***	0.76***
	(0.04)	(0.06)
Number of Observations	700	382
adjusted R^2	0.02	0.01

Notes: The table reports the results of OLS regressions. The dependent variable is the whether the participant states the unfavorable belief after the chat. Column (1) analyzes individuals whose chat partner did not mention the unfavorable belief in the chat, and Column (2) analyzes individuals whose chat partner did mention the unfavorable belief in the chat. We report in parenthesis the standard errors clustered at the chat level. Stars indicate significance at the 1%, 5%, and 10% level.

Figure 3: Screenshot of Chat Window



Notes: Chat screen of Players B with exemplary text. Participants had three minutes to chat with their partner.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election, *Journal of Economic Perspectives* **31**(2): 211–236.

Ariely, D., Bracha, A. and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially, *American economic review* **99**(1): 544–555.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F. and Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization, *Proceedings of the National Academy of Sciences* **115**(37): 9216–9221.

Bakshy, E., Messing, S. and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook, *Science* **348**(6239): 1130–1132.

Barnidge, M. (2017). Exposure to political disagreement in social media versus face-to-face and anonymous online settings, *Political Communication* **34**(2): 302–321.

Barrera, O., Guriev, S., Henry, E. and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics, *Journal of Public Economics* **182**: 104123.

Bartling, B., Valero, V., Weber, R. A. and Yao, L. (2022). Public discourse and socially responsible market behavior, *Working paper*.

Bénabou, R. (2015). The economics of motivated beliefs, *Revue d'économie politique* (5): 665–685.

Bénabou, R. and Tirole, J. (2006). Belief in a just world and redistributive politics, *The Quarterly Journal of Economics* **121**(2): 699–746.

Benoît, J.-P. and Dubra, J. (2011). Apparent overconfidence, *Econometrica* **79**(5): 1591–1625.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: Are groups more "rational" players?, *Experimental Economics* **1**(1): 101–108.

Braghieri, L. (2022). Political correctness, social image, and information transmission, *Working paper*.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A. and Roth, C. (2023). Justifying dissent, *The Quarterly Journal of Economics* **138**(3): 1403–1451.

Bursztyn, L., Rao, A., Roth, C. and Yanagizawa-Drott, D. (2023). Opinions as facts, *The Review of Economic Studies* **90**(4): 1832–1864.

Cason, T. N. and Mui, V.-L. (1998). Social influence in the sequential dictator game, *Journal of Mathematical Psychology* **42**(2-3): 248–265.

Chen, D. L., Schonger, M. and Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments, *Journal of Behavioral and Experimental Finance* **9**: 88–97.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. and Starnini, M. (2021). The echo chamber effect on social media, *Proceedings of the National Academy of Sciences* **118**(9): e2023301118.

Cooper, D. J. and Kagel, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games, *American Economic Review* **95**(3): 477–509.

Di Tella, R., Perez-Truglia, R., Babino, A. and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism, *American Economic Review* **105**(11): 4316–3442.

Drobner, C. (2022). Motivated beliefs and anticipation of uncertainty resolution, *American Economic Review: Insights* **4**(1): 89–105.

Enke, B. and Graeber, T. (2019). Cognitive uncertainty, *Working Paper 26518*, National Bureau of Economic Research.

Ewers, M. and Zimmermann, F. (2015). Image and misreporting, *Journal of the European Economic Association* **13**(2): 363–380.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. and Sunde, U. (2018). Global evidence on economic preferences, *The Quarterly Journal of Economics* **133**(4): 1645–1692.

Ging-Jehli, N. R., Schneider, F. H. and Weber, R. A. (2020). On self-serving strategic beliefs, *Games and Economic Behavior* **122**: 341–353.

Golman, R. (2022). Acceptable discourse: Social norms of beliefs and opinions, *Working paper*. Available at SSRN: https://ssrn.com/abstract=4160955.

Heatherly, K. A., Lu, Y. and Lee, J. K. (2017). Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites, *New Media & Society* **19**(8): 1271–1289.

Huffman, D., Raymond, C. and Shvets, J. (2022). Persistent overconfidence and biased memory: Evidence from managers, *American Economic Review* **112**(10): 3141–75.

Kocher, M. G. and Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games, *The Economic Journal* **115**(500): 200–223.

Kogan, S., Schneider, F. H. and Weber, R. A. (2021). Self-serving biases in beliefs about collective outcomes, *Working paper*.

Levy, G. and Razin, R. (2019). Echo chambers and their effects on economic and political outcomes, *Annual Review of Economics* **11**: 303–328.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment, *American Economic Review* **111**(3): 831–70.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of "political correctness" and related phenomena, *Rationality and Society* **6**(4): 428–461.

Masser, B. and Phillips, L. (2003). "What do other people think?"—The role of prejudice and social norms in the expression of opinions against gay men, *Australian Journal of Psychology* **55**(3): 184–190.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence, *Psychological Review* **115**(2): 502.

Morris, S. (2001). Political correctness, *Journal of Political Economy* **109**(2): 231–265.

Oprea, R. and Yuksel, S. (2022). Social exchange of motivated beliefs, *Journal of the European Economic Association* **20**(2): 667–699.

Schwardmann, P., Tripodi, E. and Van der Weele, J. J. (2022). Self-persuasion: Evidence from field experiments at international debating competitions, *American Economic Review* **112**(4): 1118–1146.

Stoner, J. A. F. (1961). *A comparison of individual and group decisions involving risk*, PhD thesis, Massachusetts Institute of Technology.

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*, Princeton University Press, Princeton.

Teger, A. I. and Pruitt, D. G. (1967). Components of group risk taking, *Journal of Experimental Social Psychology* **3**(2): 189–205.

Zimmermann, F. (2020). The dynamics of motivated beliefs, *American Economic Review* **110**(2): 337–61.