Essays in Applied Microeconomics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch

die Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Sven Andreas Norbert Heuser

aus Bonn

Bonn, 2022

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Armin Falk
Zweitreferent:	Prof. Dr. Simon Jäger
Tag der mündlichen Prüfung:	23. Juni 2022

Acknowledgments

First of all, I would like to thank Armin Falk. When I met him during my bachelor studies he showed me that economics can be far more than taking derivates and thereby played a crucial part in arising my interest in economics research. I am grateful for the exceptional trust, support and guidance he provided me with through all the years. Likewise, I want to thank Simon Jäger for being a great mentor with advice in all sorts of matters throughout my PhD. Further, I am grateful to Florian Zimmermann for all the productive and inspiring discussions, and to David Huffman for supporting me with great dedication in my last year.

I am very grateful for the financial support and great infrastructure of the Bonn Graduate School of Economics, the collaborative Research Center Transregio 224, the briq Institute on Behavior and Inequality and ECONtribute: Markets & Public Policy. Specifically, I would like to thank Stefanie Sauter and Markus Antony not only for their support in research projects but also for creating a welcoming environment at briq. Further, I would like to thank Simone Jost for her patient support in administrative matters.

Doing a PhD can be hard at times. I am fortunate to have found such close friends in my PhD, specifically Laura Ehrmanntraut, Marius Kulms, Renske Stans and Paul Voss, who helped me through it and made the years lots of fun. In particular, I am thankful to Lasse Stötzer, co-author, office mate, Puzzle collaborator, best man, and, above all, exceptional friend.

Finally, I am grateful to my parents Claudia, Norbert and my brother Torsten for giving me the chance to do a PhD and their lifelong support. Last, I am deeply thankful to my wife Laura for her love and encouragement, for believing in me and always reminding me of what really matters.

Contents

List of Figures

List of Tables

Introduction

One major goal of modern societies is to establish a cooperative and cohesive community. Though most people can agree on this goal, it is less clear how it can be achieved. To be able to make informed political changes towards its realization, it is necessary to understand not only the individual components of society in depth, the individual human being, but also how they live together and interact. Though these topics are at the heart of economics and social sciences more generally, there still remain many unanswered questions due to the complexity of individual and social behavior. For example, at a societal level, why do we observe a divide of society into distinct ideological camps and what can we do about it? Would it help to bring people together to talk? Would such a measure reduce stereotypes or could it even increase polarization? Likewise, questions of fairness and morality are similarly crucial to understand as they shape social behavior and give important insights when thinking about setting social rules. What determines moral behavior? To what extent is it driven by moral stances, and to what extent by intrinsic mechanisms and biases? However, it is also crucial to explore preferences and beliefs at the individual level. Only when we understand how beliefs and preferences are formed at the very basic level, we learn more about the way individuals ultimately behave. To give an example, why do we see so many different perceptions of how just the world is, even though everyone lives in the same world?

This thesis revolves around these broad questions. It seeks to contribute to the knowledge of how societies work and their members think and behave. Being at the heart of social sciences, it combines knowledge and tools not only from economics, but also related sciences like political science, psychology, sociology, and philosophy. The thesis comprises three chapters in total, each dealing with a different aspect of the above mentioned questions.

Chapter 1: The Effects of Face-To-Face Conversations on Polarization: Evidence from a Quasi-Experiment. Chapter 1 focuses on political preferences, beliefs and intergroup aversion. It is motivated by the perception that in many countries a division into distinct ideological camps which increasingly dislike each other takes place. The chapter sheds light on the role that conversations within and across ideological camps can play in this context. Specifically, it asks: Do conversations between like-

minded individuals exacerbate political polarization whereas conversations between contrary-minded individuals reduce it?

The study examines this question by exploiting a large-scale quasi-experiment in Germany. Within the framework of the nationwide newspaper initiative 'Germany Talks' strangers were paired for unobserved in-person meetings based on their political views. We complement this field setting with own surveys before and after the meetings. We find that talking to a person with a similar political opinion leads to more extreme political views. By contrast, meeting a contrary-minded person does not affect political views. However, it reduces negative attitudes towards those with opposing political opinions and improves the perception of social cohesion. Together, the results suggest that political in-person conversations among like-minded individuals may increase polarization of views and thus widen the gap between ideological groups, while conversations among contrary-minded individuals can reduce political intergroup aversion but not polarization of views. Hence, from a policy perspective the study demonstrates that interventions like 'Germany Talks' can be an effective countermeasure against political intergroup aversion, as long as they focus on interactions across ideological groups.

Chapter 2: Moral Luck: Mechanisms, Robustness, and Prevalence. Chapter 2 examines moral behavior. In many types of decisions, individuals can influence the probabilities of good or bad outcomes by their actions, but chance still plays a role in determining final outcomes. If punishment and rewards are conditioned on such random outcomes, this violates a property of optimal incentives. It has been posited since ancient times that humans do assign punishments and rewards based on factors outside of actors' control, a tendency called "moral luck." One famous example is the case of 'drunk driving'. Driving under the influence of alcohol increases the probability of hurting a pedestrian if she crosses the street, but the presence of her depends on chance. Do humans depend their demand for punishment only on the action 'drunk driving', independent of whether a pedestrian is hit? Meaning, that those drunk drivers who hit a person are punished equally to those who did not. Or, do they (also) depend their punishment on whether someone was actually hit?

The study provides evidence on the existence, prevalence and robustness of moral luck, and on a key open question of whether moral luck is a preference or a bias. The results are from controlled online experiments that can cleanly identify moral luck, but also involve real, consequential moral choices that are a matter of life and death for a third party (a mouse). We find moral luck in punishment, and show that this is at least partly due to a bias. Our findings support a causal chain in which random outcomes lead to biased judgments and incentivized beliefs about the nature of the actor, even though they contain zero information, and this in turn causes punishments to vary with outcomes. The study also shows that the bias is strong enough to remain in the face of an intervention that encourages deliberation. The bias is prevalent, but not universal, it is unrelated to most demographics, and is

present regardless of high or low cognitive ability or education. Finally, we also find evidence that actors exhibit internalized moral luck in how they evaluate themselves based on outcomes.

Chapter 3: Self-serving Attributions in Belief Formation. Chapter 3 studies the formation of beliefs. It is motivated by the observation that successful individuals tend to claim that their prosperity is due to their own doing and not their privileges. Living in a society where apparently "everyone can make it", they were just exceptionally good. The study proposes and evaluates a mechanism of self-serving attributions that can help to explain such beliefs and narratives.

Individuals rarely receive feedback about themselves or their performance that can be attributed to one factor with certainty. Instead, feedback is often shrouded in multi-dimensional uncertainty, i.e. there are many potential causes. In order to learn from such feedback, individuals need to make attributions to these potential causes. This study explores the idea that individuals do so in a self-serving way. More precisely, it asks whether individuals attribute positive feedback disproportionally to themselves, but negative feedback disproportionally to an external factor.

We employ a two-day laboratory experiment to present causal evidence on this question. After an IQ test on the first day, subjects receive noisy feedback on their performance on day two. The feedback depends not only on the performance but also an unknown external factor, the state of the world. We then evaluate how the feedback is attributed to the own performance as opposed to the external factor. Unfortunately, the data collection was interrupted by the COVID-19 pandemic such that control conditions comprise of fewer subjects than planned. Hence, to date only preliminary results are available. These show no clear sign that individuals attribute feedback in a self-serving way. However, there is some suggestive evidence that this is, at least partly, driven by the small size of the control groups.

Taken together, this thesis uses quasi-experimental field settings in combination with large-scale surveys, laboratory and online experiments to shed light on different aspects of preferences, beliefs and views that shape individual and social behavior. I believe that studying these questions and deepening our knowledge in this field can ultimately help to improve social cohesion and community within societies by deriving adequate policies and measures. This thesis provides a starting point for further exploration on these fascinating matters.

Chapter 1

The Effects of Face-To-Face Conversations on Polarization: Evidence from a Quasi-Experiment

Joint with Lasse Stötzer

1.1 Introduction

Political polarization has grown in many countries over recent years. Societies have become increasingly divided into distinct ideological groups and animosity between these groups has risen to a high level.[1] These trends endanger social cohesion, the functioning of democracy and even labor markets (Iyengar, Lelkes, Levendusky, Malhotra, and Westwood, 2019). Therefore, understanding what causes and how to counteract them is crucial.

According to a long-standing idea, social interactions play an important, yet two-sided role. On the one hand, there are concerns that interactions between *like-minded* individuals increase polarization as they lead to mutual reconfirmation and thus more extreme views (Sunstein, 2009). On the other hand, there is hope that interactions between *contrary-minded* individuals reduce polarization as people step out of their like-minded peer group and get to know those individuals who hold opposing views and their opinions. This idea has received substantial attention in the context of echo chambers in social media (e.g., Allcott, Braghieri, Eichmeyer, and Gentzkow, 2020; Peterson, Goel, and Iyengar, 2021). However, we still lack rigorous evidence on the effects of "real" face-to-face conversations between like-minded and contrary-minded persons. Understanding these impacts is crucial, in

1. See for example Gentzkow (2016), PEW (2014), Iyengar and Westwood (2015), and Boxell, Gentzkow, and Shapiro (2020).

particular in light of the sheer amount of face-to-face conversations in daily life and their great impact on behavior, preferences and beliefs.[2]

In this paper, we study the effects of face-to-face conversations among politically like- and among politically contrary-minded individuals on different dimensions of political polarization and social cohesion: (i) ideological polarization, i.e. how extreme political views are; (ii) affective polarization, defined as the animosity towards those with opposing political views; and (iii) the general perception of social cohesion. To estimate the effects, we leverage the quasi-experimental structure of *Germany Talks*, a nationwide newspaper initiative that matches two strangers for private in-person conversations, and complement it with surveys.[3] The conversations were neither guided nor observed. This unique combination of private yet controlled interactions in the field provides an ideal setting to study the effects of in-person conversations. We measure survey outcomes one week after the conversations.

To identify the effects of having a face-to-face conversation, we exploit plausibly exogenous variation in meeting availability. After registration, an algorithm matched two participants based on their political views. Subsequently, participants received an email in which their proposed partner was introduced. As soon as one participant accepted the proposed match, the partner was notified. If both participants accepted, contact was established and they could arrange their meeting. If at least one person did not accept, contact was not established and no meeting took place. To estimate the effects of a meeting, we restrict the analysis to those participants who accepted their partner first (*first-accepters*). This circumvents self-selection into meetings as not the first-accepters themselves but their partners decide whether contact is established and a meeting can be arranged (treatment) or no contact is established and no meeting takes place (control). However, a potential concern is that the partners' decisions depend on the first-accepters. To address this issue, we exploit the fact that all information the partner had about the first-accepter when taking the decision is contained in the introductory email. Thus, controlling for the information about the first-accepter included in the email achieves conditional random assignment of the first-accepters to treatment and control group. This approach identifies the intent-to-treat (ITT) effect of a face-to-face conversation.

To distinguish between the effects of in-person conversations with like-minded and with contrary-minded partners, we consider two treatment conditions and es-

2. In particular, in-person interactions have strong effects on political preferences (e.g., Gerber and Green, 2000; Green, Gerber, and Nickerson, 2003; Pons, 2018; Kalla and Broockman, 2020) and intergroup prejudices (e.g., Pettigrew and Tropp, 2006; Broockman and Kalla, 2016; Paluck, Green, and Green, 2019).

3. 19,000 participants registered to have a meeting. Since its launch in Germany in 2017, the program *My Country Talks* has expanded worldwide. To date, there have been interventions of the same form in many countries and regions, among others the USA (*America Talks*) and Europe (*Europe Talks*). Further countries are: Austria, Belgium, Britain, Denmark, Finland, Italy, the Netherlands, Norway, Sweden, and Switzerland.

timate respective ITT effects separately. Assignment to the two conditions is determined by the partners' difference in political views that were used for the matching.[4] The like-minded treatment and control groups contain those first-accepters in the sample, who were matched with a partner with similar political views. The contrary-minded treatment and control groups are composed of those who were matched with a partner with opposing political views. Our sample comprises 775 participants with a like- and 748 participants with a contrary-minded partner.

This paper has three main results. The first set of findings considers the effect on ideological polarization, defined as the polarization of political views towards more extreme positions.[5] We find that in-person conversations with like-minded partners increase ideological polarization, while there is no effect for contrary-minded partners. We construct two ideological polarization measures that both consider how extreme the overall political opinion - defined as a vector of eleven single political attitudes - is: the first one captures extreme views in terms of *absolute* (dis-)agreement levels on the eleven policy statements. The second one measures extreme views *relative* to the average opinion of the population. The ITT effects of having a conversation with a like-minded partner are 0.161 standard deviations more absolute and 0.166 standard deviations more relative extreme answers. By contrast, deliberating with a contrary-minded person does not affect ideological polarization. When condensing the two individual measures into one overall measure by conducting a PCA, like-minded meetings increase ideological polarization by 0.195 standard deviations. The estimates for contrary-minded conversations are negative, yet small and insignificant. As a benchmark, Allcott et al. (2020) have found that a four week long deactivation of Facebook in the US reduced their index of polarization of views by 0.1 standard deviations.

Further analysis shows that the null effects for contrary-minded conversations do not hide opposing polarizing ("backlash") and depolarizing adjustments that cancel each other out. Moreover, we detect no sign that the non-adjustment is driven by avoidance of contentious topics or shorter meeting durations. Instead, disagreement on a topic increases the likelihood of discussion and the duration of contrary-minded meetings is 20% (30 minutes) longer. Thus, contrary-minded partners discuss topics on which they disagree, but do not react to this by adapting their own opinion.

Our second set of results deals with the effect on affective polarization. In contrast to the finding on ideological polarization, we find that face-to-face conversations with contrary-minded partners reduce affective polarization while meeting a person with similar views does not have any significant impact. While affective

4. Conceptually, there are two distinct treatment and control groups within the same "framework" as the non-random matching to the partner was before the (conditionally exogenous) assignment to treatment and control.

5. In some cases, the term *issue polarization* is used when investigating changes in views (e.g., Mason, 2015; Allcott et al., 2020).

polarization is usually defined as the animosity towards partisans of the opposing party, Orr and Huber (2020) show that partisan aversion mostly reflects hostility between people with different policy views, and not hostility based on partisanship per se.[6] In line with this, we measure affective polarization by considering aversion towards people who have very different policy views in the form of stereotypes and willingness to engage in personal contact. Using a principal component analysis on all stereotypes, we find a significant reduction by 0.39 standard deviations for those who met a contrary-minded partner. This is associated with a (insignificant) higher willingness to engage in personal contact with a person with opposing views of 0.146 standard deviations. In the case of a like-minded partner, there is a (insignificant) tendency towards reinforcement of stereotypes and a reduction of willingness to engage in personal contact. When summarizing the impact on all measures into one index, contrary-minded conversations reduce affective polarization by 0.352 standard deviations, while the estimates for like-minded conversations are positive yet insignificant. As a point of comparison, a recent meta-study on the effect of inter-group contact on tolerance has found a pooled estimate of 0.39 standard deviations (Paluck, Green, and Green, 2019). Additionally, Broockman and Kalla (2016) showed that a face-to-face conversation with transgender/gender non-conforming canvassers increased tolerance by 0.45 (0.3) standard deviations three days (three weeks) after the conversations.

Our third set of findings is that conversations with contrary-minded partners improve the perception of social cohesion. Having established the impacts of in-person conversations on attitudes towards contrary-minded individuals, we turn attention to whether these effects extend to the perception of all members of the society. To assess this impact, we estimate the effects on perceptions whether fellow society members are trustworthy and pro-social. The significant ITT estimates for contrary-minded partners are 0.274 and 0.245 standard deviations, respectively. Meetings with like-minded partners show a similar, albeit weaker and insignificant tendency.

Combined, the results paint a coherent picture and provide important insights about the role of in-person conversations with respect to political polarization. On the one hand, we find that meetings with like-minded partners lead to more extreme views while they do not reduce affective polarization or bolster the perception of social cohesion. These findings suggest that the geographical clustering of people who have similar views, as reported by Brown and Enos (2021) and Bishop (2009), may widen the ideological gap between political groups further.[7] On the other hand, this

6. First, Orr and Huber (2020) find that differences in policy preferences generally lead to stronger aversion than differences in partisanship. Second, when additionally providing alignment in partisanship, aversion based on policy preferences does not change much. By contrast, when providing alignment in policy preferences, aversion based on partisanship strongly declines.

7. Moreover, the tendency towards a lesser willingness to engage in personal contact with contrary-minded individuals suggests that even the unwillingness to cross that ideological gap to interact with those who have different opinions may become greater.

paper also offers a potential solution to fight this vicious polarizing circle. We show that conversations with contrary-minded partners reduce affective polarization and improve the perception of social cohesion, although they do not decrease ideological polarization. Thus, providing people with the possibility to meet a contrary-minded person can reduce hostility across ideological groups, but does not narrow the ideological gap.

This paper relates to three strands of literature. First, we contribute to research investigating the concept of echo chambers and one-sided information provision in the context of (social) media (see e.g., Gentzkow and Shapiro, 2011; Pariser, 2011; Prior, 2013; Flaxman, Goel, and Rao, 2016; Halberstam and Knight, 2016; Martin and Yurukoglu, 2017; Bail, Argyle, Brown, Bumpus, Chen, et al., 2018; Beam, Hutchens, and Hmielowski, 2018; Sunstein, 2018; Eady, Nagler, Guess, Zilinsky, and Tucker, 2019; Di Tella, Gálvez, and Schargrodsky, 2021; Peterson, Goel, and Iyengar, 2021). In a recent paper, Allcott et al. (2020) show that the deactivation of Facebook leads to a reduction of ideological, but not affective polarization. By contrast, Levy (2021) finds that exposure to counter-attitudinal news on Facebook reduces affective polarization, but does not shift political opinions. Bail et al. (2018) even find a "backlash" effect of opinions when being confronted with opposing views on social media. We contribute to this literature by extending the analysis from (social) media to in-person conversations within and across political groups.

Second, we contribute to research exploring interventions against political polarization. Most closely related, there is research on the impact of deliberative polls that gather individuals to participate in a "mini-public" for structured and moderated group deliberations (Schkade, Sunstein, and Hastie, 2007; Fishkin, Siu, Diamond, and Bradburn, 2021).[8] Further related interventions use priming of national identity (Levendusky, 2018), correction of misperceptions (Voelkel, Chu, Stagnaro, Mernyk, Redekopp, et al., 2021), meditation (Simonsson and Marks, 2020), making outparty friendships more salient (Voelkel et al., 2021) or narrative writing (Warner, Horstman, and Kearney, 2020). We advance the literature by being the first to study the impact of one-on-one in-person discussions that are not guided or observed but take place in a natural environment, which is an important feature as the way in which conversations are held matters (Kalla and Broockman, 2020). In comparison to deliberative pollings, the conversations are more similar to every-day conversations. In addition, our design enables us to compare in-person conversations among contrary- and like-minded individuals within one quasi-experimental setup.

Finally, the paper contributes to the literature investigating whether interaction reduces inter-group prejudice. This research builds up on the contact hypothesis by Allport (1954), finding extensive evidence on the power of inter-group contact for

8. More generally, these studies explore the concept of deliberative democracy. A key part of this concept is that deliberation helps to resolve conflicts. (Habermas, 1984; Gutmann and Thompson, 2009).

various types of segregation. For example, Rao (2019) and Lowe (2021) study the effect of contact between different castes in India.[9] Meta analyses by Paluck, Green, and Green (2019) and Pettigrew and Tropp (2006) find that contact generally reduces prejudice. However, none of these studies investigate the effect of *ideological* segregation. Moreover, Paluck (2016) points out that there is a scarcity of studies that use real-world interventions with adults to test the causal effect of inter-group contact.

The remainder of the paper proceeds as follows. In Section 1.2, we briefly introduce the intervention *Germany Talks* and the political situation when it took place. Section 1.3 describes the quasi-experimental setting and our sample. In Section 1.4 we present the empirical strategy. Sections 1.5, 1.6 and 1.7 report our results, before Section 1.8 concludes.

1.2 Background

This study focuses on in-person conversations that took place within the scope of the intervention *Germany Talks* in 2018. In this section, we briefly describe the political situation in Germany in 2018 and introduce the intervention *Germany Talks*.

Political Situation. In 2018, the political divide was perceived as large in Germany. With the strong increase of asylum seekers in 2015/16, the 2013 founded right-wing party "Alternative für Deutschland" (translation: Alternative for Germany) had quickly gained popularity and with 12.6% received the third highest voting share in the federal election 2017. For the first time since WWII, a party that was more right-leaning than the established parties, such as the socially conservative Christian Democratic Union or the libertarian Free Democratic Party, had entered the German parliament, leading to a perceived overall shift to the right. Likewise, similar to other countries like the US (Iyengar and Westwood, 2015), animosity between partisans was at an alarming level, even exceeding aversion based on nationality (Helbling and Jungkunz, 2020). This prompted the federal president of Germany, Frank-Walter Steinmeier, to state in his yearly Christmas address: "Wherever you look - especially on social media - we see hate; there is shouting and daily outrage. I feel that we Germans are spending less and less time talking to each other. And even less time listening to each other."

Germany Talks. *Germany Talks* was initiated by Germany's largest weekly newspaper DIE ZEIT in 2017 as a response to the contemporary political situation in Germany. The intention behind the intervention was to enable interpersonal conversations across political camps. Since its foundation, it has established itself as a yearly

9. Other studies estimating the effect of inter-group contact include Schindler and Westcott (2021), Scacco and Warren (2018), Finseraas and Kotsadam (2017), Burns, Corno, and La Ferrara (2015), Carrell, Hoekstra, and West (2015), or Boisjoly, Duncan, Kremer, Levy, and Eccles (2006).

conducted institution with thousands of people talking to each other. Although it has its roots in Germany, the *My Country Talks* program has since expanded to other regions and countries all over the world, among others the USA (*America Talks*) and Europe (*Europe Talks*). Overall, the intervention has taken place in more than 30 countries with more than 170,000 participants to date.[10]

The basic mechanism of *Germany Talks* is simple: based on their political views, participants are matched to a partner. If both partners agree to the match, contact details are exchanged and the pair can arrange a meeting. The conversations are held in private.

1.3 Setting

1.3.1 Design

We complemented the program *Germany Talks* by sending out a baseline and an endline survey to all participants. See Figure 3.2.2 for an overview of the experimental design. The subsequent details in this section track the timeline carefully.

Recruitment. In 2018 *Germany Talks* was conducted in cooperation with a broad set of German news outlets. Together, the participating partners had considerable outreach ranging from large daily and weekly newspapers and their online platforms, over pure online media to major public television. With respect to political orientation, the participating news outlets reflected a broad political spectrum with a focus around the center-left.[11] The intervention was promoted on these platforms and participants could register either online on the respective websites or by post. 19,365 participants were successfully recruited. As shown in Figure 1.2, they came from all over Germany.

Registration. In order to register for the program, individuals had to answer seven binary political questions. Table 1.D.1 lists all seven questions, henceforth referred to as *political registration questions*. These political registration questions were chosen carefully by the organizers to capture contemporary political controversies. In addition to these questions, applicants had to state their name, age, gender, place of residence and answer five non-political free response questions.[12]

10. The program has been honored with several public awards, e.g. the Jean Monnet Prize for European Integration and the Grimme Online Award. More information can be found on https://www.mycountrytalks.org.

11. The organizing news outlet DIE ZEIT is considered as center-left. Generally, the main German media are perceived around the middle of a left-right spectrum(PEW, 2018).

12. The five free response questions were about the participants, their hobbies and dislikes. See Table 1.D.2.

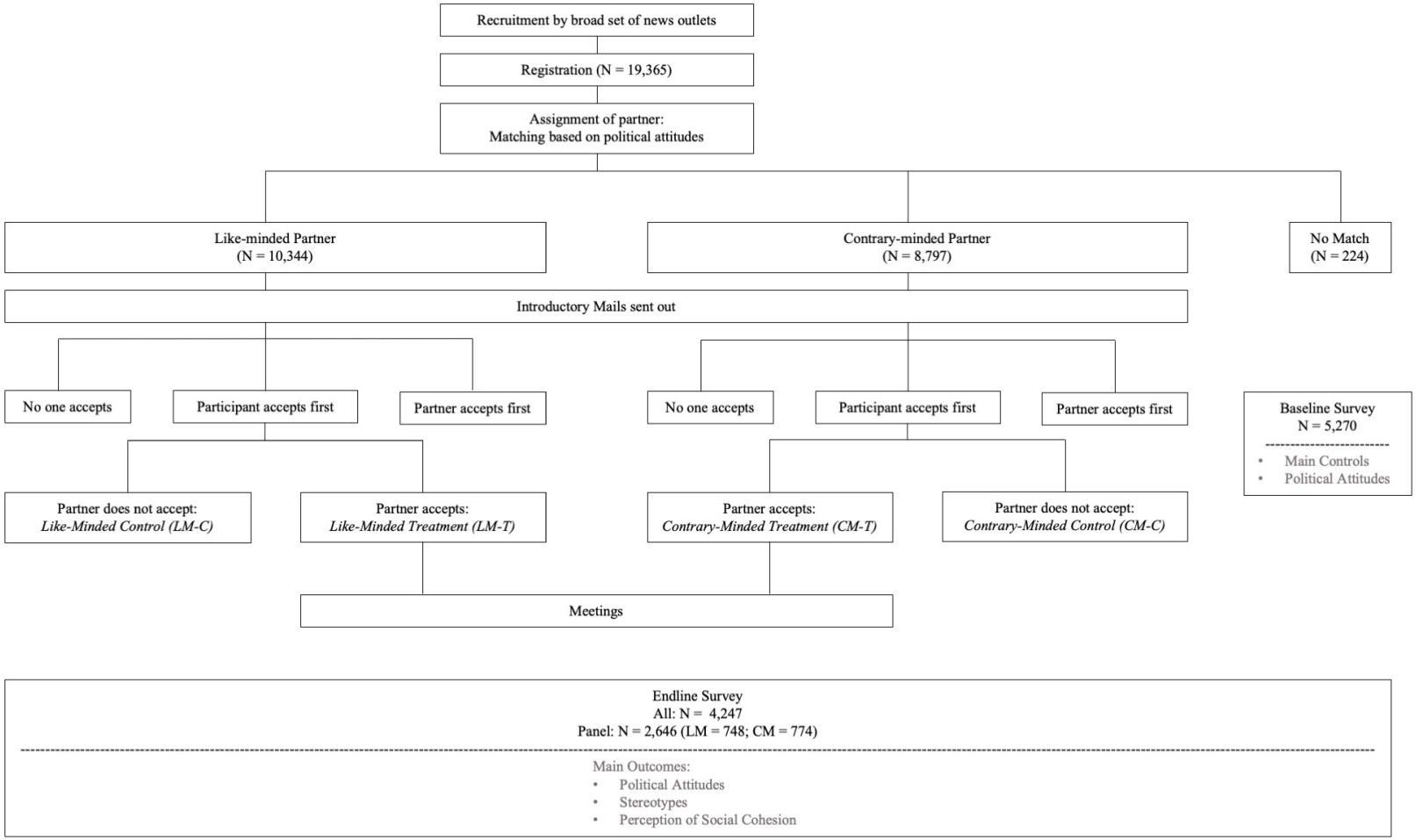


Figure 1.1. Quasi-experimental Setting

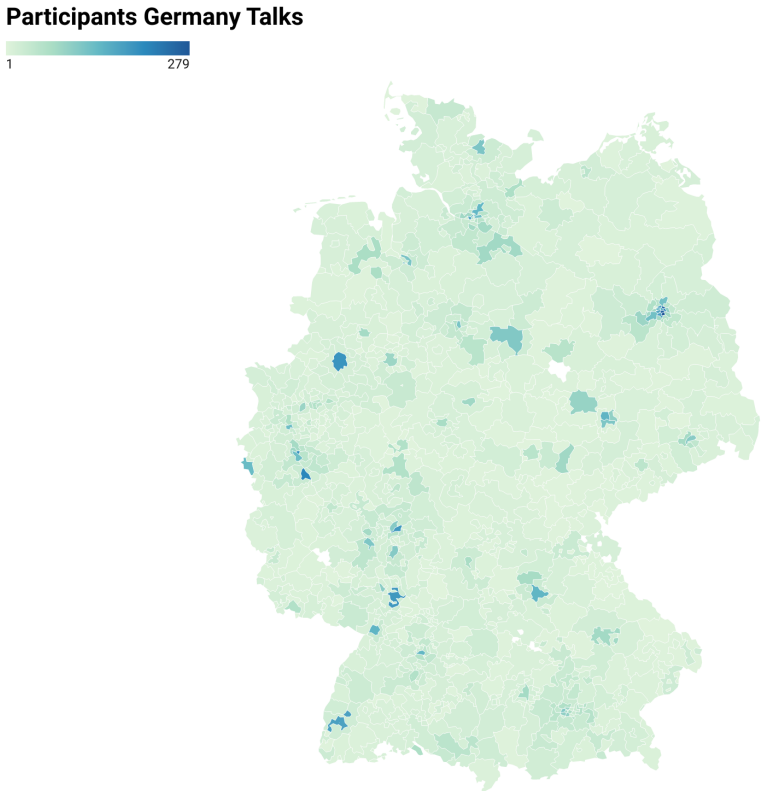


Figure 1.2. Registrations Germany Talks

Notes: Map of Germany showing the places where participants registered for *Germany Talks*. Level of visualization are NUTS regions. Blank areas depict NUTS regions where no participant registered.

Variation in Political Distance: Assignment of Treatment Condition. After registration, people were assigned a partner based on their political views and place of residence. The main objective of the algorithm was to match as many participants as possible, while fulfilling the following two conditions: First, the matched partner had to be located in a 20 kilometer perimeter.Given the fulfillment of the first condition, the *political distance* between the partners, defined as the number of differently-answered political registration questions, was maximized. The algorithm was executed exactly one time. Thus, there was no chance of changing partners or being matched to another partner later on.

We divide participants into two treatment conditions based on political distance to their partner. (i) *Contrary-Minded Partners (CM)*: This group includes those participants who were matched with a partner who had answered more than half (i.e., four or more) of the political registration questions differently. It comprises 46% of all matched participants. (ii) *Like-Minded Partners (LM)*: This group includes partic-

ipants who were matched with a partner who answered less than half (i.e., three or fewer) of the political registration questions differently. It includes 54% of the matched participants.[13]

Variation in Meeting Availability: Assignment to Treatment and Control. Each successfully paired individual received an email introducing the matched partner. This email contained a list of the political registration questions the partner had answered differently, the partner's first name, age, gender and the answers to the non-political free response questions. Based on this information, the participants could decide whether they wanted to accept the suggested partner or not. As soon as one participant within a pair accepted, the remaining partner was notified. If and only if both partners confirmed the match, contact was established by giving out the respective email addresses.

Leveraging this structure, we restrict our analysis to those participants who accepted their partner *first*, before the partner did. This leads to the fact that the (second) partner, who had not (yet) accepted, essentially decided whether the first-accepter was going to have a meeting or not. We exploit this feature by defining treatment and control groups in the following way. *Treated participants* are those first-accepters whose partner also accepted. In such cases, contact was established and the partners could arrange their meeting. *Control participants* are those first-accepters whose partner did not accept. In this case, no contact was established and there was no chance of meeting or communicating with the partner. Table 1.1 summarizes the four resulting combinations of treatment conditions LM and CM (like- vs. contrary-minded partner) and meeting availability (treatment group vs. control group).

Table 1.1. Overview Treatment & Control Groups

	Like-minded Partner (LM)	Contrary-minded Partner (CM)
Treatment (Meeting)	First-accepters, assigned to a like-minded partner who accepted as well.	First-accepters, assigned to a contrary-minded partner who accepted as well.
Control (No Meeting)	First-accepters, assigned to a like-minded partner who did not accept.	First-accepters, assigned to a contrary-minded partner who did not accept.

Notes: This table summarizes the different treatment and control groups. Treatment conditions LM and CM are shown in columns, while the rows differentiate between whether the first-accepters could arrange a meeting or not. Section 1.3 describes the assignment to treatment and control groups in detail.

13. Throughout the paper, we show that the results are robust to alternative sample splits into like- and contrary-minded partners.

There are two key points for this paper. First, rather than first-accepters selecting themselves into the treatment and control group, the partners of the first-accepters assign the first-accepters to the treatment and control group. Second, the partners could base their decision on whether to also accept or not merely on the information about the first-accepters from the introductory email. Thus, conditional on that information, the decision was independent of the first-accepter.

Meetings. After contact had been established, the organizers of *Germany Talks* played no further role and participants had to organize the exact time and location of the meetings themselves. Meetings were not observed, nor moderated or guided in any way. They mostly took place in natural settings like cafes, parks, or in people's homes. As shown in Figure 1.3, conversations centered around the topics of the seven political registration questions. On average, conversations lasted 140 minutes and an overwhelming majority of the participants reported that it was a pleasant experience.

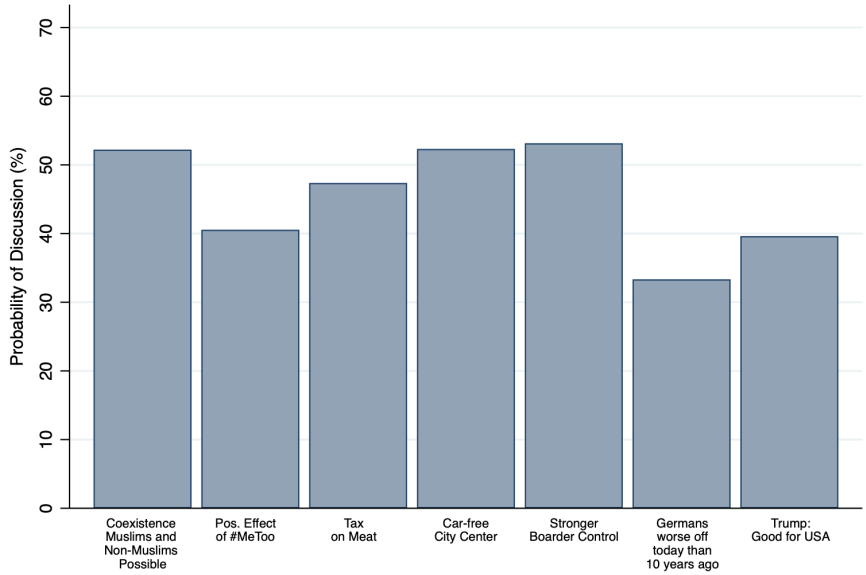


Figure 1.3. Topics of the Conversations

Notes: This figure plots the probabilities of discussion for the seven political registration questions. The y-axis of the graph denotes the frequency in %. Table 1.D.1 shows the political registration questions.

Surveys. Baseline and endline surveys were sent out by the organizers of *Germany Talks*. Unfortunately, the baseline survey was distributed more than one week after the introductory emails had been sent. Therefore, first-accepters' assignments to the partner (treatment condition), acceptance decisions and assignments to treatment

(acceptance decision of the partner) had already taken place before most participants filled out the baseline survey. In fact, by that point in time 98% of the treated participants had already learned that the partner had also accepted.[14] Consequently, measures that were elicited in the baseline survey may potentially be affected by first email contact between partners or expectations. For this reason, we only use measures from the baseline survey that are robust.[15]

Basic information about the participants like socio-demographics was only elicited in the baseline survey. It was sent out five days prior to the meetings and required on average 14 minutes to answer. Besides the outcome measures, the endline survey contained questions about the meetings, if they had taken place. The average response time was 12.5 minutes. It was sent out one week after the conversations. 2,645 participants completed both surveys. Additional details on the surveys can be found in Appendix.

1.3.2 Sample

In our study, we focus on first-accepters who filled out both surveys. Table 1.2 describes the composition of the resulting sample, which comprises 1,523 participants. Compared to the German population (column 1), our sample (column 2) is similar in terms of age, income and place of residence, but more educated, male, politically left-leaning and with less migration background.[16] While the sample is left-leaning *on average*, it is not clear how this translates to the existence of distinct ideological groups within the sample. Are all participants from one "left" political camp, or are there still a left and right group represented in the sample? As shown in Table 1.2, party preferences and self-classified ideology suggest the existence of a large left camp and a small right camp. To further explore this heterogeneity, we investigate correlational patterns of the answers to the political registration questions. The organizers of *Germany Talks* carefully picked them in a way that there is typically a more "left" and a more "right" answer.[17] Thus, we should expect that one group

14. Participants had time to accept until the day when the meetings took place. Thus, in principle, first-accepters had the chance of becoming a member of the treatment group until that moment.

15. In particular, we do not use any sensitive "social measures" like stereotypes or perception of social cohesion. We solely utilize time-invariant measures and political attitudes.

16. There are two potential reasons for these differences. On the one hand, different types of people may differ in their willingness to participate in a program promoting political discussion. For example, conservatives may be less willing to have such a discussion. This case may be partly seen as a feature of our study as voluntary participation - in contrast to "forced" or paid interpersonal conversations - is an important requirement for the success of such policies in real life. On the other hand, the specificity of the sample may also reflect the reader-/viewership of the participating news outlets. We cannot clearly differentiate which of the two factors plays how much of a role, but it is likely to be a mixture of both.

17. There are questions like "Should Germany increase its border control?", which represent typical left vs right topics, in this case migration. Other questions, like "Is Donald Trump good for the USA?"

gathers around left answers while another group chooses predominantly right answers, if there are actually members of the two distinct camps within our sample. To check this, we use latent class analysis.[18] LCA endogenously creates classes with specific answer patterns and assigns each participant a likelihood of membership in each class. Applying it to all registered participants, we see a bipolar distribution, i.e. participants belong to either one or the other class with a high probability (see Figure 1.C.1). Assigning participants to classes according to the probabilities, we find a large group to which 82% and a small group to which 18% of the participants belong. The answer patterns of the two groups, shown in Figure 1.C.2, confirm the hypothesized distinction into a (large) ideologically left and a (small) ideologically right group. Membership in the left group predicts agreement with more liberal notions and clear disagreement with more conservative viewpoints. Likewise, members of the right group show a rather conservative answer pattern.[19] A t-test using self-stated left-right classification confirms the interpretation with the members of the large group being significantly more left ($p<0.01$). To further support this finding, Table 1.D.3 reassuringly shows that we find nearly identical groups if we use k-means clustering instead of LCA. Focussing on the sample that we use, it is representative of all registered participants in terms of class membership (83% and 17%). Taking all facts together, our sample comprises a majority of left- and a minority of right-leaning participants.

Table 1.2. Summary Statistics

	German Population (%)	Sample (%)		
		All	LM	CM
Age				
18 - 34	24	25	27	23
35 - 54	32	37	35	39
55 or older	43	38	37	39

do reflect less classic left-right topics, but nevertheless yield predictions about what conservatives and liberals would answer.

18. LCA is related to factor analysis as both explore the relationship among variables. However, in contrast to FA, LCA assumes a categorical latent variable with a multinomial distribution instead of a continuous normal-distributed variable. This method does not demand any a priori assumptions about the correlations between the questions (i.e. which answers should belong in which group). Instead, it takes the data and checks whether there are latent classes whose members have specific answer patterns.

19. For example, membership in the left group predicts disagreement with the demand of stricter border control, and agreement with the notion that #metoo had some positive effects. Membership in the right group predicts agreement with stricter border control, but shows otherwise a less differentiating pattern. This is unsurprising as many of the conservative answer options are rather extreme opinions. For example, disagreement with the statement that the #metoo movement and the debate about sexual harassment had *some* positive effects arguably reflects a far right position.

Table 1.2. (continued)

	German Population	Sample		
		All	LM	CM
Gender				
Female	49	37	42	32
State				
Baden Württemberg	13	13	13	14
Bayern	16	14	14	14
Berlin	4	13	16	11
Brandenburg	3	2	2	3
Bremen	1	1	1	0
Hamburg	2	6	7	5
Hessen	8	8	8	9
Mecklenburg-Vorpommern	2	1	0	2
Niedersachsen	10	10	11	9
Nordrhein-Westfalen	22	17	16	18
Rheinland-Pfalz	5	3	3	3
Saarland	1	1	1	1
Sachsen	5	5	5	5
Sachsen-Anhalt	3	1	1	1
Schleswig-Holstein	3	4	4	3
Thüringen	3	1	0	2
Migration background				
Yes	23	10	10	10
Education				
No Education	2	0	0	0
Lower Sec. Education	24	1	1	1
Middle School	30	7	6	7
Advanced technical certificate	6	6	7	6
High School	10	17	17	17
University	27	67	68	66
Other	0	1	1	2
Income (monthly; EUR)				
0-800	19	10	11	8
800-1499	25	13	13	13
1500-2199	23	20	21	20
2200-3299	17	23	26	21
3300 or more	17	27	24	30
Political spectrum left-right				
Far-left	3	4	4	3
Left	18	25	29	21
Centre-left	30	40	44	34
Centre	28	20	18	21
Centre-right	16	9	4	15
Right	3	2	0	4

Table 1.2. (continued)

	German Population	Sample		
		All	LM	CM
Far right	1	1	0	1
Party				
Die Linke	10	14	14	12
Bündnis/90 Die Grüne	16	50	54	39
SPD	17	11	12	9
CDU/CSU	28	7	5	8
FDP	9	7	5	9
AfD	15	7	0	13
Other	5	5	3	5
Don't Vote/Don't know	31	2	1	2
Ideological Class				
Left Ideology		83	98	67
Right Ideology		17	2	33
Observations		1,523	775	748

Notes: This table presents characteristics of the German adult population, our sample, and the like-minded (LM) and contrary-minded (CM) subsamples. Measures for the German population are taken from the German Microcensus (age, gender, marital status), German Allbus 2018 (education, migration background, income, religious confession, religiousness), the CSES 2017 (left-right), and an election poll by Forsa from the week prior to DS (Party). To allow for comparisons, some variables were transformed by collapsing several subcategories into one supercategory.

Subsamples. Columns 3 and 4 of Table 1.2 provide descriptive statistics of the subsamples of like-minded and contrary-minded first-accepters. Subsample sizes are similar with 775 participants in the like-minded and 748 in the contrary-minded condition. The two subsamples are comparable, except for political preferences, with the like-minded sample being less conservative. The reason for these political differences lies in the mechanics of *Germany talks*: with a large part of the registered participants being from the left ideological camp and the matching algorithm aiming to maximize political distance between partners, conservatives were predominantly matched with left participants. Analogously, liberals often ended up being matched with fellow liberals due to excess supply. Consequently, the like-minded subsample contains left but no right people, while the contrary-minded subsample comprises left and right people.

1.3.3 Treatment Conditions: Like- and Contrary-Minded Partners

The treatment conditions differ in the political views of the partners who are by construction like- or contrary-minded to the first-accepters. Table 1.D.4 provides descriptive statistics of the partners. It shows that in the like-minded condition they

are younger, more female and more left than in the contrary-minded condition, as would be expected following the rationale about pair compositions above.

To assess the extent to which the treatment conditions actually reflect politically like- and contrary-mindedness within pairs, we compare them with an alternative way of defining of whether a person met a like- or contrary-minded partner. As each participant of *Germany Talks* can be assigned one ideological class found by the LCA, this allows us to use the overlap of ideological classes within pairs to define like- and contrary-mindedness. As shown in Table 1.D.4, there is strong congruence of our treatment conditions and the overlap of ideological classes within pairs. This gives further substantial foundation to our treatment condition definitions. For robustness, we also report results using the overlap of ideological classes to define treatment conditions.

1.4 Empirical Strategy

Specification. Our approach identifies the ITT of having an in-person conversation with either a like- or contrary-minded person. Recall that the partner assigns the participant who accepted the match first (first-accepter) to treatment and control by choosing to accept or not, based only on the information from the introductory email. Thus, by controlling for the information from the introductory emails the assignment is conditional independent of the first-accepter. While we are able control for most of the content from the mails, we have to use proxies for the surname and the answers to the open questions from the participants.[20]

For both treatments LM and CM separately, we estimate the following ITT specification by OLS:

$$Y_i = \alpha + \beta * Treat_i + \gamma * BasicInfo_i + \delta * AddInfo_i + \rho * Y_i^b + \epsilon_i \qquad (1.1)$$

where Y_i denotes our outcome variable from the endline survey. The dummy $Treat_i$ indicates whether first-accepter i was accepted by the partner or not and ϵ_i is an individual-specific error term. β measures the intent-to-treat effect of a political face-to-face discussion. $BasicInfo_i$ and $AddInfo_i$ are sets of fixed effects capturing the information from the introductory mails, and Y_i^b denotes the baseline value of Y.[21] $BasicInfo_i$ contains basic information (hard facts) about participant i that we observe (age intervals, gender, region at the NUTS level, combinations of answers to political

20. We know age, gender, answers to the political registration questions, and region. Due to data protection, we did not receive surname nor the answers to the open questions from the organizers of *Germany Talks*.

21. Y_i^b excludes the baseline values for the measures of affective polarization and perception of social cohesion as treatment conditions had already been assigned and contact had already been established in almost all cases when baseline values were elicited. For more details see Section 3.2.

registration questions) and proxies for surname (migration background, and education and income). The set of dummies $AddInfo_i$ accounts for the fact that the answers to the open questions were unobserved by capturing potentially visible information. It comprises political self-classification (left to right), party, political engagement, religion, religiousness, marital status and the number of politically contrary-minded people in one's social environment. Appendix 1.B.1 describes the controls in more detail.

The main identifying assumption is that we achieve conditional independence of treatment assignment and the respective outcome variable by controlling for $BasicInfo_i$ and $AddInfo_i$. This would be violated if, for example, some attitudes of the participants shine through in the introductory mail, consequently affect the partners' decisions, and importantly also have an impact on the outcome variable.

For robustness, we also report estimates from OLS regressions without $AddInfo_i$ and for the post-double-selection (PDS) method (Belloni, Chernozhukov, and Hansen, 2014). Out of the vector of all potential controls, PDS chooses the right set via a three-step "double-lasso" procedure: using two lasso regressions, it selects a set of controls that is predictive of treatment status $Treat_i$ and a set of controls that predicts outcome Y_i. In a third step, the union of both sets of control variables is used to estimate the treatment effect. The conclusions from all three specifications are the same. If anything, the PDS method yields smaller standard errors and thus "more significant" estimates.

Potential Challenges. Table 1.3 suggests conditional random assignment to the treatment and control groups in both conditions LM and CM is achieved. None of the coefficients that are not affected by the treatment are significant in one of the treatment conditions LM and CM, nor is the F-Test of joint significance. Table 1.D.5 shows that the treatment and control groups are even conditionally balanced if we use the more conservative approach of conditioning only on the basic set of controls.

Table 1.D.6 tests for conditional selective attrition between the baseline and endline survey. Note that income (part of the basic controls *BasicInfo*) and marital status (part of the additional controls *AddInfo*) are not controlled for because we (only) elicited them in the endline survey. Thus, we should interpret the findings with caution. We find very small and insignificant differences between the treatment and control groups in both the LM (column 1) and CM (column 2) conditions. Mean attrition is 49% in both cases.

As many participants already knew their treatment status before the baseline survey was sent, people may have selected differently into our panel depending on the treatment condition. Table 1.D.7 tests for selective response rates to both surveys between the treatment conditions. Note that, as none of controls from the surveys can be used (because the surveys are part of the test), assignment to treatment and control is not conditionally exogenous. Thus, the findings are only suggestive and should be interpreted cautiously. There are significant, yet small differences

Table 1.3. Balance Checks

	Like-minded Partner (1)		Contrary-minded Partner (2)	
Political Views				
Border Control	0.0969	(0.137)	-0.0922	(0.139)
#metoo	-0.191	(0.127)	-0.103	(0.148)
Meat Tax	0.00334	(0.140)	-0.0752	(0.189)
Car free inner-cities	-0.163	(0.132)	-0.0806	(0.158)
Coexistence (Non-)Muslims	-0.0415	(0.114)	0.0486	(0.149)
Germans worse off	-0.00698	(0.157)	0.0500	(0.169)
Trump	-0.0387	(0.0981)	0.0764	(0.126)
Same-sex marriage	-0.118	(0.122)	-0.161	(0.153)
Cooperation within EU	-0.114	(0.0973)	0.172	(0.122)
Income Tax	0.118	(0.160)	-0.0373	(0.172)
Trustworthiness Media	0.0310	(0.160)	-0.148	(0.169)
Importance				
Border Control	0.0357	(0.222)	0.219	(0.232)
#metoo	0.0737	(0.178)	-0.152	(0.204)
Meat Tax	-0.0495	(0.177)	0.150	(0.196)
Car free inner-cities	0.0474	(0.178)	0.184	(0.192)
Coexistence (Non-)Muslims	0.161	(0.157)	0.0729	(0.172)
Germans worse off	0.326	(0.216)	0.182	(0.222)
Trump	0.285	(0.224)	0.186	(0.235)
Beliefs				
Number applications for asylum	-16641.0	(33678.8)	-8822.1	(41681.3)
Share Muslims in Population	-0.177	(0.601)	0.107	(0.741)
F-Test	0.95		0.71	
P-Value	0.52		0.82	

Notes: The table reports the treatment coefficients of the balance checks. Dependent variables are measures from the baseline survey: baseline political views, subjective evaluation of importance of political topics, and baseline beliefs about the share of muslims in Germany and number of asylum seekers in Germany. Each of these variables is regressed on the treatment dummy and the sets of basic and additional controls. The respective dependent variable is listed in the left column. Column (1) reports the results for the like-minded and column (2) for the contrary-minded individuals. F-Tests of joint significance are calculated by regressing the treatment on all those variables and the sets of basic and additional controls. Robust standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

between the treatment and control groups in both treatment conditions (6.7% and 7.2%). 18.9 and 21.5% of all participants fill out both surveys in the LM and CM condition, respectively.

To assess to the extent to which the intent-to-treat effect captures the real effect of a face-to-face meeting, we look at compliance with treatment assignments. Since contact was only established if both partners had accepted, by construction non-compliance is only one-sided. Participants in the control group had no chance to

meet their partner.[22] Compliance with treatment status is very similar across both treatment conditions, at 87.2% for LM and 86.8% for CM. Thus, the high compliance rates of 100% (control) and 87% (treatment) suggest that the average effects of the meetings are close to our ITT estimates. They are presumably even slightly larger, as the ITT likely provides a lower bound with some participants in the treatment group not having a meeting.

One potential challenge to the interpretation of our study is that we estimate the effects separately in two subsamples of different (political) compositions. Differences in effects may partly be rooted in the differences between subsamples instead of being caused by the treatments.[23] To assess the extent of the concern, we look at the selection into the different subsamples in more detail. Table 1.D.8 shows that we do not see any signs that the willingness to accept the partner first varied with political distance. Thus, together with the discussion on subsample differences from the previous section, it seems that the subsamples are in large parts comparable except for political orientation (see Table 1.2). To account for the observed differences in political attitudes, we re-weight our contrary-minded sample to match the like-minded sample means using the entropy weighting procedure (Hainmueller, 2012). We find the same pattern, which suggests that it is unlikely that the differences in effects are only found due to the dissimilarity of the subsamples.

1.5 Effects on Ideological Polarization

Many scholars argue that deliberations among citizens lead to more agreement within society. However, there is the concern that discussions can yield the exact opposite. Like-minded people may confirm and reinforce each other's opinion (Sunstein, 2009) leading to more polarized views. Even if confronted with contrasting viewpoints, it is unclear what to expect as discussions may result in a "backfire" effect (Wojcieszak, 2011; Bail et al., 2018). In this section, we therefore explore the heterogeneity in effects of interpersonal deliberation on political opinion.

Measures. To measure polarization in political opinions, we elicited agreement with eleven different political viewpoints in the baseline and endline survey. See Table 1.4 for an overview. Seven out of the eleven viewpoints were those used by *Germany Talks* to match partners.The remaining four viewpoints capture other typical left-right topics, such as same-sex marriage. We define the overall political opinion as the vector of all eleven opinions. We construct two measures that each capture one

22. There were two participants who stated that they met a partner even though the partner did not accept them. We do not know whether they lied on purpose or accidentally stated that they met their partner. We drop them from our analysis, but including them in our analysis does not change our results.

23. Note that this does not concern the identification of the ITT of like- vs contrary-minded meetings.

Table 1.4. Outcome Variables

Variable	Statement
Political Views	
Overall Political Opinion	
Coexistence	Muslims and Non-Muslims can coexist in Germany.
#metoo	The public debate about sexual harassment and #metoo had some positive effects.
Tax Meat	Meat should be taxed higher in order to reduce its consumption.
Car-free City Centers	German city centers should be car-free.
Border Control	Germany should implement stricter border controls.
Germans worse off	Germans are worse off today than 10 years ago.
Trump	Donald Trump is good for the USA.
Same-Sex Marriage	Marriage should only be allowed between a man and a woman.
Cooperation within EU	Germany should deepen its cooperation with other EU countries.
Income Tax	To reduce the gap between rich and poor, the tax rate for top earners should be increased.
Trustworthiness Media	Altogether, German media are trustworthy.
Affective Polarization	
Overall Stereotype	
Cognitive Abilities	This person is incapable of understanding complex contexts.(rev.)
Poorly Informed	This person is poorly informed.
Moral Values	This person has completely different moral values.
Way of Life	This person leads a completely different life.
Willingness to Engage in Personal Contact	I would like this person to be in my personal environment.(rev.)
Perception of Social Cohesion	
Trustworthiness	One can trust most people in Germany.
Pro-Sociality	Most people in Germany do not care about the wellbeing of others.

Notes: The table shows all elicited variables that we use to construct our outcome measures. *Overall Political Opinion* is a vector consisting of the eleven single political views. Out of this vector we construct both ideological polarization measures. See Section 1.5 for more details. *Overall Stereotype* is the first principal component of a PCA of all four stereotypes as detailed in Section 1.6. To elicit the affective polarization measures, we asked participants to picture some person that gave *very different* answers to the seven political attitude questions. The last column shows the corresponding scales. Some variables, denoted by (rev.), are reversed for interpretational reasons. Participants had to state their agreement to the statements (political attitudes, perception of social cohesion) and the extent to which they apply (stereotypes) on seven-point Likert-Scales.

facet of ideological polarization. The first measure captures how extreme the overall opinion is in terms of *absolute* (dis-)agreement with the viewpoints. More precisely, it is defined as the Euclidean distance to the center of the scale. The second measure captures how extreme the overall opinion is *relative* to the average opinion of the population. Put differently, it reflects the extent to which the opinion is aligned with the average opinion of the population. It is constructed in an analogous way to the first measure and is defined as the Euclidean distance to the average pre-meeting opinion of the subsample. To estimate the overall effect on ideological polarization, we condense the two individual ideological polarization measures into one measure via principal component analysis. Using one measure yields effect sizes that usefully summarize the overall impact of the conversations on ideological polarization and allows us to benchmark effect sizes. All outcome measures are standardized by subtracting the respective control group means and dividing by the control standard

deviations. For more information on construction of the outcome measures, see Appendix 1.B.

Findings. Figure 1.4 presents ITT effects for the two individual and the overall ideological polarization measures. It shows that the conversations significantly polarized those participants who met a like-minded partner but not those who met a contrary-minded partner. The ITT effects on the two individual measures are 0.161 and 0.166 standard deviations in the like-minded treatment condition, respectively. The point estimate of the overall effect being 0.195 standard deviations is slightly larger than in the case of the two individual measures. For those who met a contrary-minded partner all point estimates are negative, yet insignificant. In particular, we do not find any sign of backlash effects. Figure 1.C.3 shows the ITT effects for the post double selection method (PDS). The figure confirms the findings. The point estimates are similar. However, the estimates are more precise, as the number of controls is much smaller, yielding more narrow confidence intervals.

Tables 1.D.9 and 1.D.10 provide the respective estimation results for the whole set of controls, the post double selection method (PDS) and a smaller set of controls. The results are very similar across specifications. Tables 1.D.11 and 1.D.12 test whether results are robust to an alternative treatment condition definition based on membership to the ideological classes found by the latent class analysis: instead of defining whether a person met a like- or a contrary-minded person by using the number of different answers to the partner, this approach uses the alignment of class memberships of the partners. The results do not change. Table 1.D.13 confirms the findings if treatment condition definitions are varied by splitting participants into like- and contrary-minded based on alternative cut-offs: participants are assigned to the like-minded condition if they coincide with their partner in three or more and five or more political registration questions, respectively (instead of four or more). The definition of the contrary-minded treatment condition is varied analogously. Tables 1.D.14, 1.D.15, 1.D.16 and 1.D.17 provide the results when using alternative distances measures, Manhattan distance and Mahalanobis distance to construct our variables instead of Euclidean distance. We find largely the same pattern. Table 1.D.18 tests whether results change when like-minded regressions are reweighted to match contrary-minded means in political preferences (party affiliation, self-reported left-right classification), gender and age. Likewise, contrary-minded regressions are reweighted to match the like-minded sample. Results are very similar suggesting that the differences between like-minded and contrary-minded effects are not only found because their different (political) compositions.

One potential reason for the null effect in the contrary-minded condition is that it masks heterogeneity as found in other persuasion studies (Baysan, 2021). In this case, polarizing (backfire) and de-polarizing (intended) effects would cancel each other out. This may happen for different attitudes within one person, or, alternatively, for different persons. To shed light on this, we look at the general overall

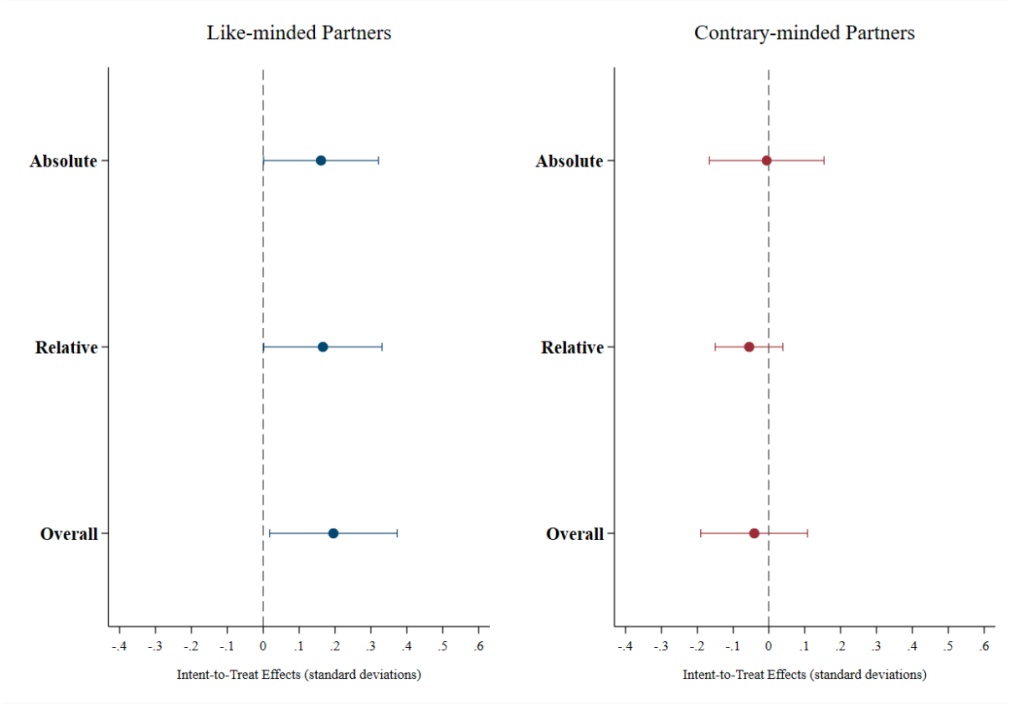


Figure 1.4. Effect of the Conversations on Ideological Polarization

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on the three standardized measures of ideological polarization. It plots the effects on how extreme the overall political opinion is (i) in terms of absolute (dis-)agreement to policy views, and (ii) in relation to the average opinion of the population. (iii) It plots the effect on the overall measure of ideological polarization, defined as the first principal component of the two individual measures. Higher values are associated with more ideologically polarized (extreme) outcomes. The outcome measures are described in Section 1.5 and regression specifications are detailed in Section 1.4. 95% confidence intervals are included.

change defined by the mere Euclidean distance between the base- and endline political opinion. This measure focuses on the amount of change and ignores its "direction". Figure 1.C.4 plots the corresponding ITT effects and shows that in general only conversations with like-minded partners lead to a substantial adjustment of one's own political opinion.

Why is there no adjustment in contrary-minded conversations? The findings by Chen and Rohla (2018), who show that Thanksgiving dinners are significantly shorter when residents from opposing-party precincts attend, suggests that participants may avoid contentious topics. In contrast to this hypothesis, the meetings among contrary-minded partners were significantly longer than those among like-minded partners, with median durations of 150 and 120 minutes, respectively ($p < 0.01$). Figure 1.5 plots the probabilities that contrary-minded partners talked about a specific topic depending on whether a pair agreed or disagreed on it. The graph

shows that disagreement clearly increases the likelihood of discussing a particular topic. The results suggest that the effects are not driven by the avoidance of topics between contrary-minded persons. By contrast, participants particularly discuss contentious topics and learn about their partner's viewpoint, but do not alter their own opinion due to it.

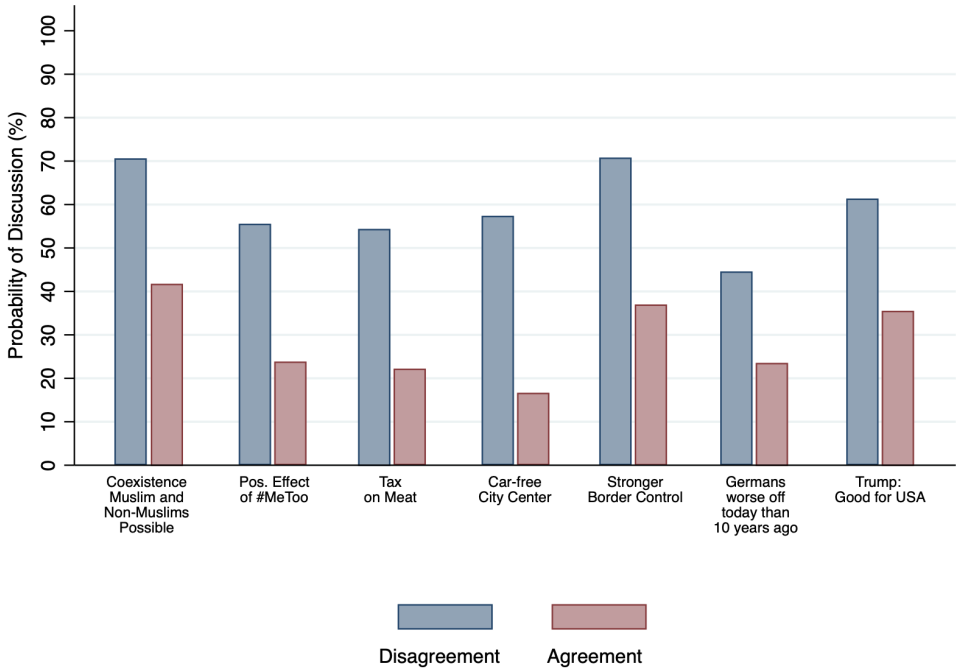


Figure 1.5. Conversational Topics: Agreement vs Disagreement (CM)

Notes: The figure plots probabilities of discussion for the seven political registration questions in the contrary-minded treatment condition, depending on whether the partners agreed or disagreed on the topic. The Y-axis indicates the share of pairs that discussed the respective topic. Table 1.D.1 shows the political registration questions.

Is the ITT effect for like-minded meetings large? As one benchmark, we can compare the overall effect size to those of related interventions. Allcott et al. (2020) study the impact of a four week long deactivation of Facebook on political polarization in the US. They find a reduction in their index of issue polarization of approximately 0.1 standard deviations. Our overall effect size is nearly twice as large. Further, we can follow Allcott et al. (2020) and set our estimates in relation to the change in a different index of several political polarization measures in the US (Boxell, 2020). The author finds an increase of 0.38 standard deviations between 1996

and 2016. With 0.195 of a standard deviation, our ITT estimates is about 50 percent of that increase.[24]

1.6 Effects on Affective Polarization

Beyond the effect on ideological polarization, political discussions may have an impact on affective polarization. Independent of the change of their political opinion, people may adjust their view about those who have different opinions. Indeed, related research on prejudice reduction through interaction suggests that interpersonal conversations between contrary-minded persons may lead to a reduction of stereotypes (Allport, 1954; Kalla and Broockman, 2020; Fishkin, Siu, Diamond, and Bradburn, no date). In this section, we therefore turn attention to estimating the impact of face-to-face discussions with members of one's own and the other political camp on affective polarization.

Measures. To assess the effect on affective polarization, we use two measures, namely stereotypes about and preference for personal contact with contrary-minded persons. We defined such contrary-minded persons as someone who has opposing political views on the seven political registration questions.[25] We elicited stereotypes about contrary-minded persons that were communicated by former participants of *Germany Talks*. These were the prejudices that contrary-minded individuals are cognitively less capable, poorly informed, have different moral values and lead completely different lives. We reduce dimensionality by implementing a principal component analysis (PCA). We use the first principal component which is the convex combination of the four stereotypes that accounts for the largest possible variation in the data, as our overall stereotype measure. Table 1.D.21 provides the respective loadings (weights). To gain a broader picture, we additionally measured the preference for close interpersonal contact with opposing political views. More precisely, we elicited participants' willingness to have a contrary-minded person in their social environment. See Table 1.4 for a detailed overview of the outcome measures.

Stereotypes. Figure 1.6 shows that interpersonal conversations with contrary-minded persons significantly reduced stereotypes. The point estimate is -0.379 standard deviations. Figure 1.C.6 estimates the ITT effects on each stereotype separately. The reduction is strongest for the belief that contrary-minded persons are of low

24. Of course, these benchmarking exercises need to be interpreted with caution: For example, the samples of our study are very different from those by Allcott et al. (2020) and (Boxell, 2020). In particular, both papers look at US residents while our study took place in Germany. Furthermore, the measures of issue and political polarization of Allcott et al. (2020) and (Boxell, 2020) differ from our measure of ideological polarization.

25. Note that we did not elicit beliefs and attitudes towards the partner, but towards some arbitrary person with opposing views.

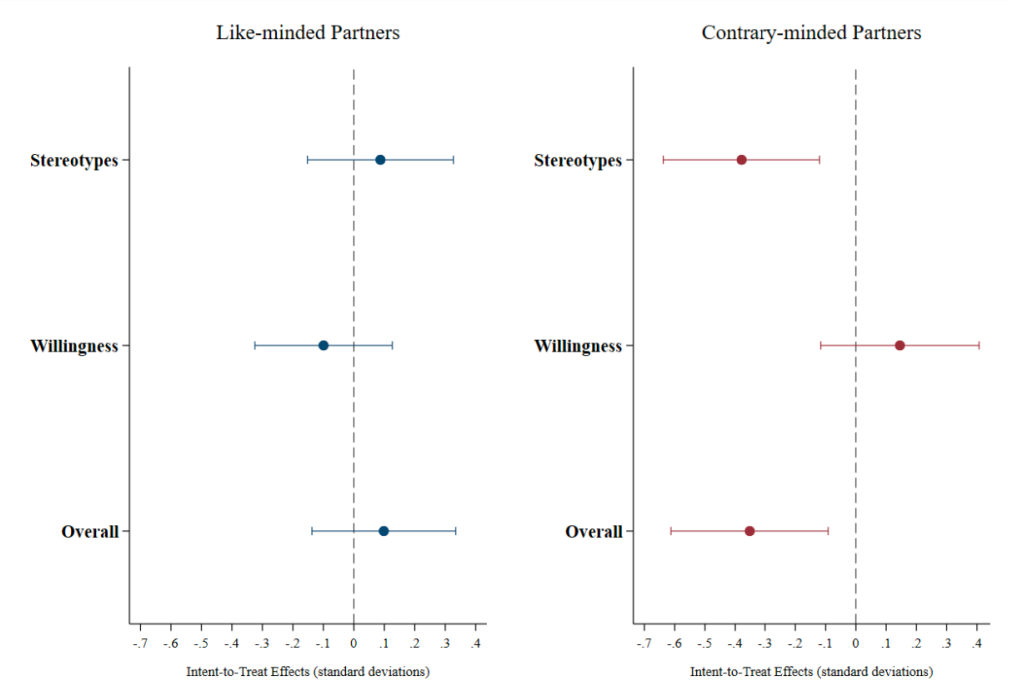


Figure 1.6. Effect of the Conversations on Affective Polarization

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on (i) standardized overall stereotypes about a person with opposing political views, (ii) standardized willingness to engage in personal contact with a person that has opposing political views, and (iii) standardized overall affective polarization. The overall stereotype measure is defined as the first principal component of all four elicited stereotypes. Table 1.D.21 shows the loadings. Lower values denote lower stereotypes (and lower affective polarization). Lower willingness to engage in personal contact is associated with higher affective polarization. The overall affective polarization measure is defined as the first principal component of all four elicited stereotypes and the willingness to engage in personal contact. Table 1.D.29 shows the respective loadings. Lower values are associated with lower affective polarization. The measures are described in Section 1.6 and regression specifications are detailed in Section 1.4. 95% confidence intervals are included.

cognitive ability, while we do not see any decrease in whether contrary-minded persons lead a completely different life. Meeting a person from one's own political camp does not have any effect on stereotypes about contrary-minded persons. The positive point estimate of 0.087 standard deviations suggests that if anything conversations with like-minded partners tend to slightly increase stereotypes. However, none of the effects is significant, for neither the overall nor for the individual stereotypes. Figure 1.C.5 plots the ITT effects for the post double selection method (PDS) and confirms the findings. The point estimates are slightly smaller, yet more precise.

Tables 1.D.20, 1.D.24, 1.D.25, 1.D.23 and 1.D.22 show the robustness of the results to dropping controls, and running PDS regressions for the overall and indi-

vidual stereotypes. Tables 1.D.13 and 1.D.26 show that the effects are similar if treatment conditions definitions are altered by varying the cut-off and using alignment of ideological classes, respectively. Table 1.D.27 provides the results when like-minded regressions are reweighted to match the contrary-minded sample, and vice-versa. We find the same pattern.[26]

Willingness to Engage in Personal Contact. Figure 1.6 presents the effect of the conversation on willingness to engage in personal contact with a contrary-minded person. In line with the previous finding, the point estimate for meetings with a contrary-minded partner is 0.146 of a standard deviation meaning a stronger willingness to engage in personal contact, yet insignificant. Analogously, the coefficient for like-minded meetings is -0.0993 and insignificant. Figure 1.C.5 shows the effects for the post double selection method (PDS). The estimate of contrary-minded conversations is of a similar size (0.176 standard deviations) but significant at the 5% level due to a smaller standard error. Similarly, the coefficient for like-minded partner is -0.137 standard deviations and significant at the 10% level. Table 1.5 shows the respective estimates and robustness to dropping the set of additional controls (columns 1 and 4). Varying the definition of like- and contrary-minded partners produces very similar results (see Tables 1.D.13 and 1.D.28). Table 1.D.27 shows robustness towards reweighting the subsamples.

Interpretation. The results for stereotypes and willingness to engage in personal contact paint a coherent picture. To estimate the overall effect on affective polarization, we conduct a PCA with all five affective polarization measures, the four stereotypes and willingness to engage in personal contact. Hence, the resulting overall measure is a weighted index of the five measures capturing aversion towards contrary-minded persons.[27] This usefully summarizes the overall impact on affective polarization and allows benchmarking effect sizes. Figure 1.6 provides ITT estimates for both treatment conditions. The estimates for like-minded partners are insignificant, but positive (0.099 standard deviations), while conversations with contrary-minded persons reduce affective polarization by 0.352 standard deviations ($p<0.01$).[28]

To put the effect magnitude in perspective, we use two different benchmarks. First, we follow Lowe (2021) and compare our estimates with effects of inter-group contact from a recent meta-analysis by Paluck, Green, and Green (2019). The meta-analytic effect of 0.39 standard deviations is very close to our estimate. Second,

26. The effect sizes of like-minded meetings are even slightly larger. This suggests that the effect may partly be driven by left leaning individuals.

27. Table 1.D.29 provides the loadings on the overall measure. With positive signs for the individual stereotypes and a negative sign for willingness to engage in personal contact, it confirms the interpretation of an overall measure for animosity towards contrary-minded persons.

28. Figure 1.C.5 shows the effects for PDS.

Table 1.5. Effect on Willingness to Engage in Personal Contact

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-0.113	-0.0993	-0.137*	0.131	0.146	0.176**
	(0.110)	(0.115)	(0.0799)	(0.122)	(0.133)	(0.0779)
Constant	0.733	-0.563		1.149	0.211	
	(1.196)	(1.104)		(0.991)	(1.482)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	755	755	755	727	727	727
R^2	0.394	0.501		0.529	0.582	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized willingness to engage in personal contact. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): Various combinations of the political registration questions, various NUTS FE. The specifications are described in more detail in Section 1.4.
* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Broockman and Kalla (2016) show that a ten-minute face-to-face conversation with transgender/gender non-conforming canvassers leads to an increase in tolerance. The effect sizes are 0.45 standard deviations after three days and 0.3 standard deviations after three weeks, respectively. Our effect consistently ranks between both the two points in time of elicitation (the endline survey being sent out seven days after the conversations took place), and the two effect sizes. The fact that Broockman and Kalla (2016) found very long lasting effects after a ten-minute conversation may give hope that our conversations with a median duration of 150 minutes lastingly reduced affective polarization.

1.7 Effects on the Perception of Social Cohesion

One fear associated with the rising levels of affective and ideological polarization is the threat to society as a whole (Iyengar, Lelkes, et al., 2019). The increasing gaps and animosity between contrary-minded individuals may threaten social cohe-

sion by changing how society members are perceived. Although the contact hypothesis predicts improved attitudes towards contrary-minded persons, it is less clear whether these effects also transfer to general levels of beliefs and attitudes. Related evidence by Rao (2019) finds an increase of general pro-sociality after contact, while Lowe (2021) observes a reduction of general trust.[29] In this section, we hence shed light on the effect of interpersonal conversations on perceptions of trustworthiness and pro-sociality of fellow society members.

To explore the heterogeneous impact of interpersonal conversations, we elicited two beliefs: first, the belief about how trustworthy fellow citizens generally are, and second, the belief about to the extent to which German citizens generally care about the well-being of others (see Table 1.4).

Findings. Figure 1.7 provides the ITT effects on the two beliefs. For both types of conversations, the point estimates are positive for both measures, although in the case of like-minded conversations they are small and insignificant. Coefficients for contrary-minded meetings are 0.274 (trustworthiness) and 0.245 (pro-sociality) standard deviations and significant.

Tables 1.D.30 and 1.D.31 provide estimates for the PDS regressions and if the set of additional controls is dropped. The results are similar, although the PDS effect on trustworthiness for meetings between like-minded partners is also significant due to a slightly larger coefficient and smaller standard error. Tables 1.D.13, 1.D.32 and 1.D.33 show the robustness of the results towards varying the definition of treatment conditions. Table 1.D.34 provides reweighted results and finds largely the same pattern.

To assess the overall impact of the conversations on the perception of social cohesion, we summarize both perceptions into one measure by using a PCA. Figure 1.7 plots the corresponding ITT effects. In line with the effects on the individual measures, the estimate for contrary-minded meetings is 0.299 standard deviations. The like-minded coefficient is positive, yet insignificant.

The findings are in large parts in line with the effects on affective polarization and the idea that the positive inter-group effects extend to attitudes towards a more general population. Conversations among contrary-minded individuals reduce affective polarization and have a positive impact on the perceptions of general trustworthiness and pro-sociality. However, the (insignificant) tendencies for like-minded conversations are not consistent with the hypothesis. Although affective polarization tends to increase, trust and perception of general pro-sociality both also tend to improve.

Alternative Explanation: Disappointment. One potential alternative explanation of our findings on affective polarization and social cohesion may be that disap-

29. Similarly, Dinesen, Schaeffer, and Sønderskov (2020) show that ethnic diversity is generally negatively related to generalized trust.

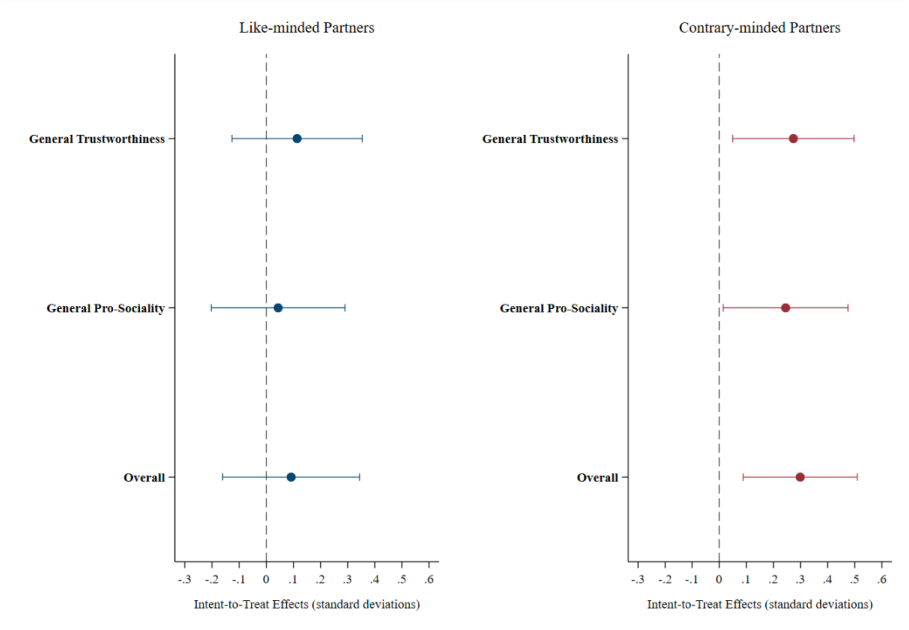


Figure 1.7. Effect of the Conversations on the Perception of Social Cohesion

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on standardized measures of perceptions of social cohesion. It plots the impacts on (i) the perception that fellow citizens are generally trustworthy, (ii) the perception to what extent fellow citizens generally care about the well-being of other and (iii) the overall effect, defined as the first principal component of the two primer measures. Higher values denote higher perceptions. The outcome measures are described in Section 1.7 and regression specifications are detailed in Section 1.4. Error bars reflect 95% confidence intervals.

pointment of not being accepted by the proposed contrary-minded partner drives the effects. To assess this concern, we compare the time trends of the two control groups. If disappointment with not being accepted by the contrary-minded partner is actually increasing affective polarization, we should see different time trends for the contrary- and the like-minded control group as the latter were not rejected by contrary-minded partners. Table 1.D.35 finds no sign for different time trends.[30] This suggest that disappointment does not explain the effects for affective polarization and perception of social cohesion for contrary-minded partners.

30. Note that the comparison makes use of the baseline data, which we carefully avoided in our analysis. Even though the concern may be smaller when comparing participants who did not have contact with their partner prior to the baseline survey, the results should be interpreted carefully.

1.8 Conclusion

This study exploits a natural experiment to estimate the impact of political face-to-face conversations on political polarization. It provides evidence that in-person communication among people who hold similar political views further fortifies these opinions. As a consequence, existing differences in opinions between different political camps are magnified, making people even more unequal in their opinion how policy should be shaped. One could argue that differences in policy views are not negative by themselves given that a healthy democracy "is designed" to handle such disagreements. However, as soon as people condition their attitudes and behavior on other people's political opinions, this argument begins to fall apart. In this respect, the paper provides evidence that communication across political camps can help. It shows that talking to someone who holds contrasting political views reduces negative attitudes towards contrary-minded persons and improves the perception of social cohesion. Therefore, the study provides clear policy implications. It shows that reducing obstacles to communicating with contrary-minded people and facilitating interaction between different political camps can be an effective countermeasure against affective polarization. One possibility to achieve this may be interventions like "My Country Talks". However, these interventions should focus on interactions between groups. More generally, our findings support any effort to bring together to talk those who hold different views. People may understand each other better without having to give up their own convictions.

This study explores the effects of one single in-person conversation. It therefore provides a benchmark for the possible effects of echo chambers. At the same time, it serves as a proof of concept that, given the right circumstances, interpersonal communication is a powerful tool.

One limitation of this study is that due to the quasi-experimental contraints, it does not explore long-term effects on polarization. Further, it would be interesting to explore whether the observed effects are also reflected in behavioral changes. Another weakness is rooted in the nature of our sample being a selection of people who want to deliberate on politics. The impact of conversations, in particular with contrary-minded persons, may differ for those who have a lower willingness to do so. However, from a policy perspective, the sample at hand may be the right one to look at as these types of persons can actually be reached via relatively simple policies.

Appendix 1.A Additional Details on Germany Talks and Surveys

1.A.1 Media, Recruitment and Meetings

Participating Media. These news outlets were DIE ZEIT, Süddeutsche Zeitung and SZ.de, tagesschau.de and Tagesthemen (ARD aktuell), Deutsche Presse-Agentur, Der Spiegel, Chrismon and evangelisch.de, Schwäbische Zeitung, Die Südwest-Presse, Der Tagesspiegel, t-online.de, and Landeszeitung Lüneburg. The majority of the news outlets are traditional print media with online appearances. For example, DIE ZEIT is the largest weekly newspaper and Süddeutsche Zeitung is the second-largest daily newspaper in Germany. Both also cover the Internet and Broadcast Media. t-online.de is a pure online news outlet. Tagesthemen is a daily news show in the evening on ARD, one of the two major German public television channels. On 16/08/2018 Tagesthemen showed a clip inviting viewers to participate in the program.[31] tagesschau.de is the online appearance of ARD. According to PEW (2018), ARD is the main news source for many Germans. This holds for people across the political spectrum. The political orientation of the larger partners is center/center-left. PEW (2018) show that ARD, Der Spiegel, and Süddeutsche Zeitung are placed on the middle of the left-right spectrum. Freitag, Kerkhof, and Münster (2021) measure the political position of news outlets by politicians' sharing behavior. They conclude that DIE ZEIT and Der Spiegel are positioned on the left of the political spectrum. ARD and Süddeutsche Zeitung are positioned on the center-left.

Registration Process. Participants were recruited by the news outlets. They could register online on the respective websites and additionally via mail (DIE ZEIT). To register the participants had to answer the *political registration questions*, seven Yes or No questions about contemporary political topics that were chosen by the program organizers of *Germany Talks* to be as controversial as possible.[32] The translated questions can be found in Table 1.D.1. After answering the political registration questions, individuals were introduced to the program. They were told that if they choose to participate, the program would attempt to find a person residing within a 20 km radius from their home who answered the seven questions differently and is willing to meet at a predetermined date (September 23, 2018). If an individual decided to participate, the email address, zip code, name, gender, and age of the individual were collected, as were the answers to five questions in which participants were asked to describe themselves. The five questions are listed in Table 1.D.2.

31. The clip is available under following link (in German): Link.

32. The whole intervention was designed by the organizers of *Germany Talks*. We took no part in designing the intervention.

Meetings. Participants had to organize the exact time and location of the meetings themselves. However, the suggested and officially communicated date of the conversations was September 23, 2018. 90% of the participants reported to have met on that date. The meetings were unobserved: There was no third-party moderating, guiding, or observing the discussion and no rules or topics of discussion were predefined. On average the conversations took 2 hours and 20 minutes. The shortest reported meeting was 40 minutes, while the longest meeting was 10 hours. These numbers indicate the participants took time to get to know the other person and discuss their (opposing) viewpoints.

To shed light on what happened during the meetings, we elicited the topics of the conversations and details about about the atmosphere during the conversation and the general experience of being part of *Germany Talks*. Figure 1.3 plots how frequent the topics of the *political registration questions* were discussed. These topics are at the core of our political attitude measures. We see that the conversations centered around these topics. The least discussed topic of the *political registration questions* was whether Germans are worse off today than 10 years ago (33%). The most discussed topics were: Stronger border control (53%) and car-free inner cities (52%). Moreover, if a pair disagreed on a topic, the likelihood of discussing it is higher than in the case of agreement. Figure 1.5 plots the likelihoods of discussion if the partner agreed and disagreed for contrary-minded pairs. Overall, the meetings were a pleasant experience: 95% of the participants stated that the atmosphere during the conversation was enjoyable, 94% said that there were no loud or heavy disputes and 75% stated that their conversation partner was likable.[33]

1.A.2 Surveys

As a complement to the program *Germany Talks*, we designed two surveys. The surveys were sent out by the organizers of *Germany Talks*. One survey was sent out prior to the suggested and officially communicated date of the conversations (baseline survey) and one after the conversations took place (endline survey).

Baseline Survey. All registered participants were invited to fill out the baseline survey. The baseline survey was sent out five days before the suggested day for the conversations (18/09/2018). At this point, the email introducing the matched partner had been out for a week and 98% of the treated participants had already learned that the partner had accepted. 5,677 participants took the survey. The average response time was 14 minutes. The elicited measures are described in detail in Appendix 1.B.

33. Participants had to state how much a statement applied to their conversation on a seven-point Likert-Scale. The reported percentages are for those who stated one of the two highest categories, *agree* or *strongly agree*.

Endline Survey. All registered participants were invited to participate in the endline survey. The endline survey was sent out eight days after the conversation (01/10/2018). Even though the organizers of *Germany Talks* strongly suggested holding the conversation on 24/09/2018, not all participants were able to meet on the specified day. However, 97% of the respondents had met at least 3 days before we sent out the email. 4,200 participants completed the survey. The average response time was 12.5 minutes. The elicited measures are described in detail in Appendix 1.B. Out of the 4,200 responders, 63% also answered the baseline survey.

Appendix 1.B Measures

Our analysis relies on two datasets: data from the intervention *Germany Talks* and self-reported survey data. The primary dataset consists of all 19,134 registered participants and includes age, gender, zip-code, answers to the seven political registration questions and the matched participant. The latter dataset consists of information elicited in the baseline or the endline survey. We have all data points for 2,465 participants.

1.B.1 Controls

In our analysis we condition on a variety of control dummies that stem from both datasets, the *Germany Talks* and the survey dataset. In the baseline survey, we gathered information about participants' demographics like education, migration background, and religion, the political heterogeneity of their social environments, i.e. how many politically contrary-minded people they have in their social environment, and their political preferences, which includes a position on a political self-classification and the party they would vote for. In the endline survey we elicited income and marital status. The following paragraphs list the relevant controls and how we construct them.

Set of Basic Info. The set of dummies *BasicInfo* contains basic information (hard facts) about the participant that we observe (age intervals, gender, region on NUTS level, combinations of answers to political registration questions) and proxies for surname (migration background, and education and income). More precisely, we divide age into following six intervals: 18-25, 26-35, 36-45, 46-55, 56-65, 65+. Gender is a binary variable indicating whether a person identifies as male, female or nonbinary. Instead of including 1531 five-digit zip codes in our analysis, we construct dummies based on the Nomenclature of Territorial Units for Statistics (NUTS) to increase power. NUTS (level 3) is a geocode standard that is developed and regulated by the European Union and divides Germany into 401 regions. We include all combinations of the seven binary political registration questions to control for policy view patterns. From our baseline survey, we include variables for the participants' education, income, and migration background. Education is an ordinal variable with seven categories from "No school leaving certificate" to "Ph.D.". We include dummies for each category. Migration background is a binary dummy, where we define a person with a migration background as someone who either was not born in Germany or has parents who were born in a different country. Income is an ordinal variable that captures the net income per month of the respondents. It contains five categories, from "0-800 Euro" to "3300+ EUR" and an option for participants that don't know their monthly income. All variables additionally have a category "Not specified".

Set of Additional Info. The set of dummies *AddInfo* accounts for the fact that the answers to the open questions were unobserved by capturing potentially visible information. We did not receive that information (and the surname) by the organizers of *Germany Talks* due to data protection. Thus, we use proxies to capture potential topics as well as possible. Table 1.D.2 shows the five open questions. *AddInfo* consists of *dummies* for each category of the measures party preference, political self-classification, political engagement, religion, religiousness, marital status, and the number of politically contrary-minded people in their social environment. Party preference indicates the party that the respondents would vote for. It is a nominal variable with nine categories including all five parties represented in the 19th Bundestag (German parliament) and the categories "Other party", "I don't know", and "I do not vote". Political self-classification is an ordinal variable with seven values from "Very liberal" to "Very conservative". Political engagement contains different forms of political engagement that participants have been part of or not: "Participation in civic initiatives", "Attending demonstrations", "Being an active member of a party", and "Being an active member of a trade union". Religion is a nominal variable indicating religious affiliation (7 categories). Religiousness is an ordinal variable eliciting how often participants visit a place of worship. It has six categories from "Never" to "More than once per week". Marital status dummies are "Single", "Divorced", "Widowed", "Registered partnership", "Married and living separately", "Married and living with a spouse". The number of contrary-minded people in the participants' social environment contains seven categories from "None" to "All". For all variables, we add a dummy indicating a missing value.

1.B.2 Outcome Measures

Outcome measures were elicited in the endline survey. Only in the case of political views, we also use values from the baseline survey to construct our measures. All outcome measures are standardized by subtracting the (respective) control group mean and dividing by the control group standard deviation.

Political Views. Participants were asked to state the extent to which they agree with different political statements on a seven-point Likert scale. Apart from the transformation from questions into statements and the change of scales, the first seven of the eleven statements were identical to the political registration questions. In addition to the seven questions, we elicited four other, more general political attitudes. See Table 1.4 for an overview. Based on these attitudes, we create outcome measures for our analysis. The underlying idea is to take all eleven attitudes together and interpret the eleven-dimensional vector as the overall political opinion. In contrast to the measures of affective polarization and perception of social cohesion, we use data from the baseline survey as political views are not as easily affected by either learning the treatment condition (like- or contrary-minded partner) or first email contact with the partner. Importantly, looking at individual

changes enables us to do a more precise analysis.

Change towards More Extreme Views: Absolute (Dis-)Agreement We construct two measures of ideological polarization. The first measure indicates to what extent a person shows stronger (dis-)agreement to the topics after the meeting. More precisely, we construct one measure that indicates whether someone moved towards or away from the midpoint of our scale (a vector of 3s), denoting neither disagreement nor agreement. The measure is defined as follows:

$$ExtremeViewsAbsolute_i = \sqrt{\sum_{s=1}^{11}(Y_{si2} - 3)^2} - \sqrt{\sum_{s=1}^{11}(Y_{si1} - 3)^2}$$

where Y_{sit} denotes individual i's level of agreement to statement s in the endline (t=2) and the baseline (t=1) survey. The eleven statements are the political attitudes from Table 1.4. The first term is the Euclidean distance between i's agreement and the center point (vector of 3s) in the endline survey (t=2), while the second term is the respective Euclidean distance in the baseline survey (t=1). Thus, $ExtremeViewsAbsolute_i$ indicates the change in the distance to the midpoint of our scale. A positive realization of this variable indicates that individual i moved "towards the boundary of our scale", whereas a negative realization implies that i's attitudes changed "in the direction of the center". If the variable equals zero, participants moved neither closer nor further away from the center.

Change towards More Extreme Views: Relative to Population The second measure of ideological polarization reflects the change in the extent to which an individual's overall opinion aligns with the average overall opinion in the respective subsample (treatment condition):

$$ExtremeViewsRelative_i = \sqrt{\sum_{s=1}^{11}(Y_{si2} - \overline{Y}_{s1c})^2} - \sqrt{\sum_{s=1}^{11}(Y_{si1} - \overline{Y}_{s1c})^2}$$

where Y_{sit} denotes individual i's level of agreement to statement s in the endline (t=2) and the baseline (t=1) survey. The eleven statements are the political attitudes from Table 1.4. $\overline{Y}_{s1c}$ is the average level of agreement to statement s of all participants in the treatment condition c in the baseline survey. The two terms reflect the distance to the average pre-meeting opinion after and before the meeting took place. In sum, $ExtremeViewsRelative_i$ denotes whether someone moved towards ($ExtremeViewsRelative_i < 0$) or away from ($ExtremeViewsRelative_i > 0$) the average pre-meeting opinion or none of the two.

General Change of Political Opinion To measure the general adjustment of the political opinion we construct a measure that disregards any direction, but focuses on the

mere amount of change. More precisely, we define general change as the Euclidean distance between end- and baseline survey:

$$GeneralChange \;=\; \sqrt{\sum_{a=1}^{11} (Y_{si2} - Y_{si1})^2}$$

where Y_{asit} denotes individual i's level of agreement to statement s in the endline (t=2) and the baseline (t=1) survey. The eleven statements are the political attitudes from Table 1.4.

Affective Polarization. To study how the conversations' affected stereotypes about individuals with contrasting political views and participants' willingness to have personal contact with these individuals, participants had to picture a person that gave opposing answers to the seven political registration questions. We then elicited participants' beliefs about this person by asking them to which extent they agree with different statements about the contrary-minded person on a seven-point Likert scale. Importantly, we did not elicit beliefs and attitudes towards the matched partner but some generic person that hold opposing views. The elicited stereotypes were communicated by previous participants of *Germany Talks*.

Stereotypes - We elicited four stereoytpes. These were the beliefs that contrary-minded persons have low cognitive abilities, are poorly informed, have different moral values and lead a different life. Table 1.4 shows the exact wordings. We condense these questions by conducting a principle component analysis. We use the first principle component as our overall *stereotype* measure. A higher value of our *Stereotypes* measure is associated with larger stereotypes about contrary-minded individuals. Table 1.D.21 provides the loadings of the first principle component.

Willingness to Engage in Personal Contact We elicited participants' *willingness to engage in personal contact* by asking participants to state their level of agreement to the statement that they do not want to have a person with opposing views in their social environment. For our analysis, we reverse the scale. See Table 1.4 for the exact wording.

Perception of Social Cohesion. To assess the effect on participants' perceptions of social cohesion in Germany, we elicited two beliefs. First, we asked how trustworthy the fellow citizens in Germany are (*Perception of General Trustworthiness*). Second, we measured participants' *Perception of General Pro-Sociality* by asking to what extent German citizens generally care about the wellbeing of others. The two questions are listed in Table 1.4.

Appendix 1.C Additional Figures

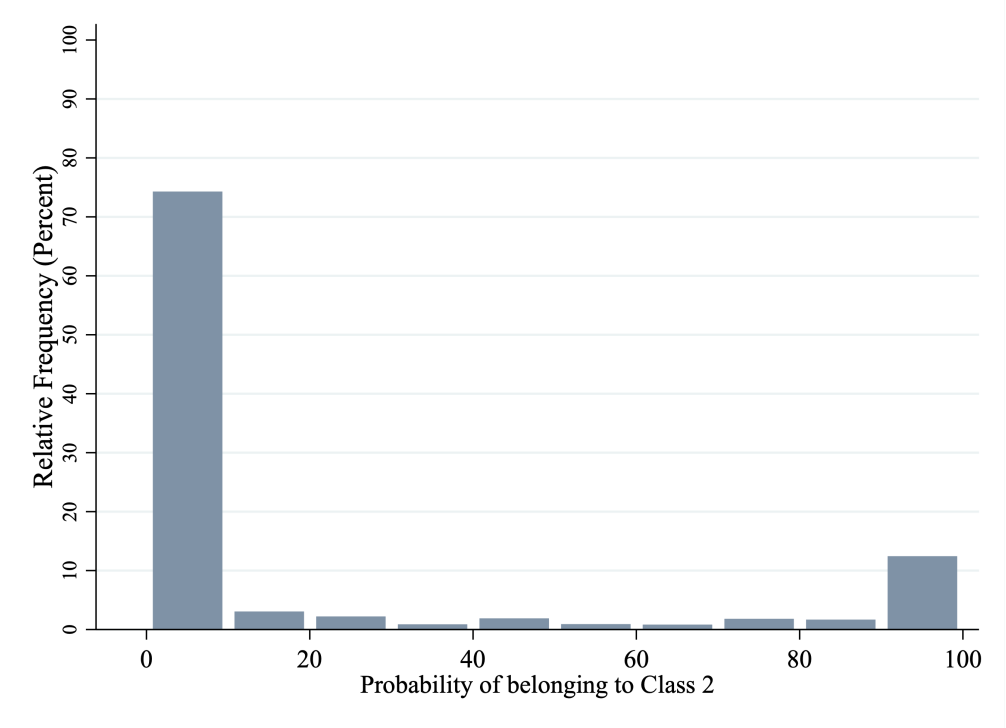


Figure 1.C.1. LCA: Likelihood of Class 1 Membership

Notes: The Figure plots the distribution of probabilities to belong to class 1 from the Latent Class Analysis. The LCA is described in Section 1.3.

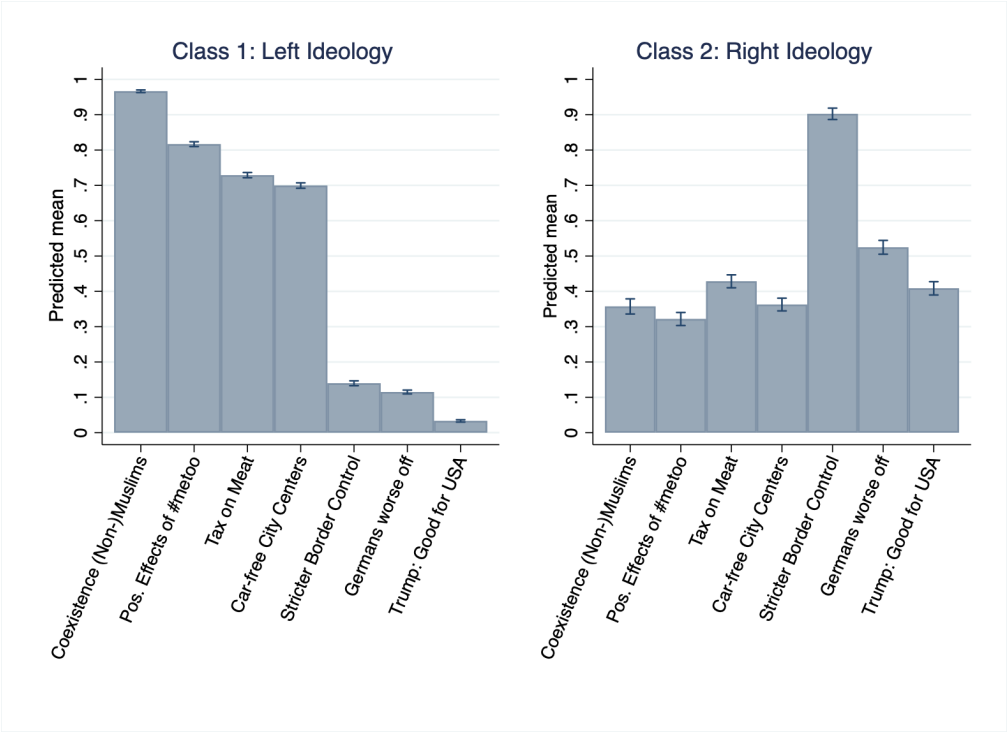


Figure 1.C.2. LCA: Conditional Likelihood of Agreement

Notes: The Figure plots the probabilities of agreeing to the binary political registration questions conditional on LCA class membership. The political registration questions are shown in Table 1.D.1 and the LCA is described in Section 1.3. Error bars reflect 95% confidence intervals.

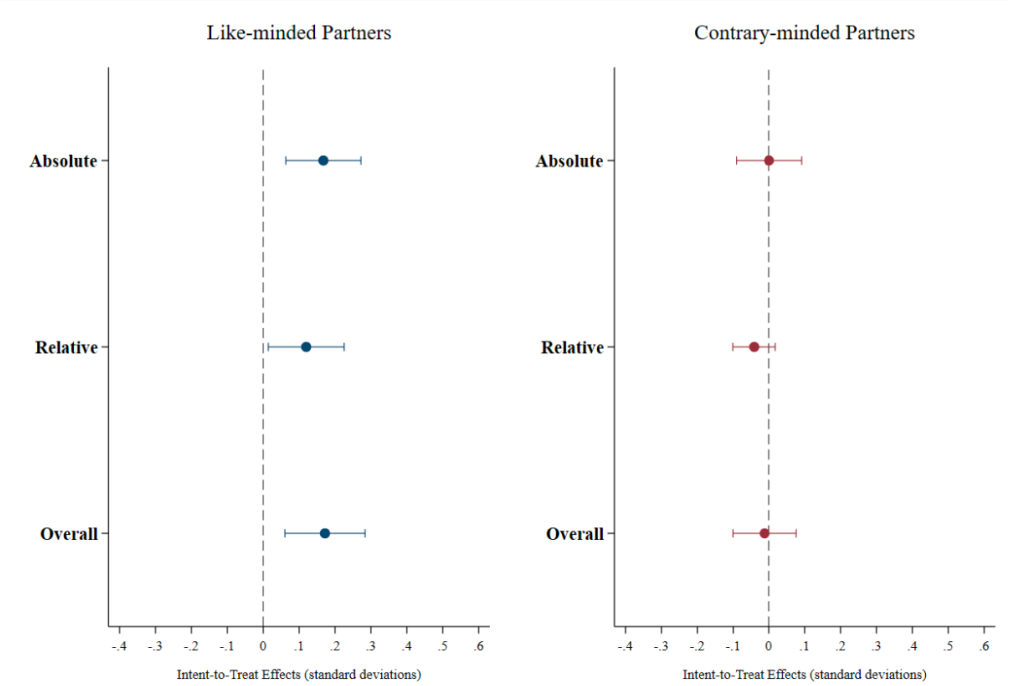


Figure 1.C.3. Effect of the Conversations on Ideological Polarization (PDS)

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on the three standardized measures of ideological polarization for the post double selection method (PDS). It plots the effects on how extreme the overall opinion is (i) in terms of absolute (dis-)agreement to policy views, and (ii) in relation to the average opinion of the population. (iii) It shows the effect on the overall measure of ideological polarization, defined as the first principal component of the two individual measures. Higher values are associated with more ideologically polarized (extreme) outcomes. The outcome measures are described in Section 1.5 and regression specifications are detailed in Section 1.4. 95% confidence intervals are included.

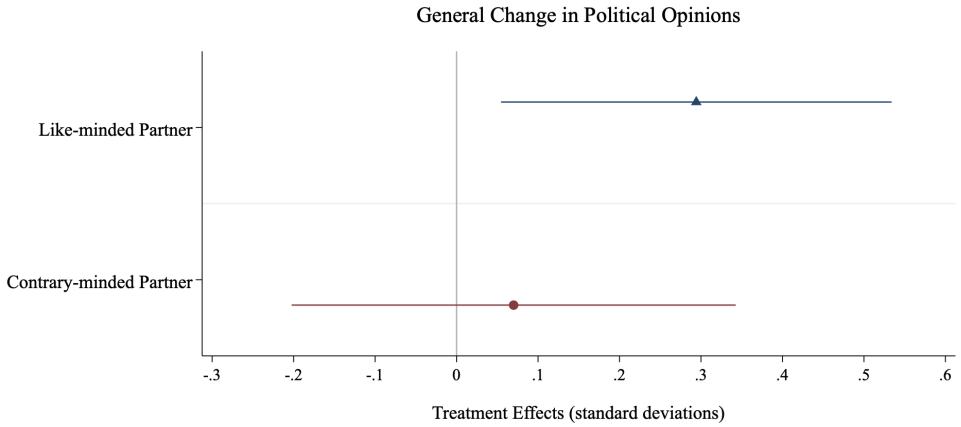


Figure 1.C.4. Effect on General Change of Political Opinion

Notes: This figure shows the ITT effects of the like- and contrary-minded treatments on standardized general change of the overall political opinion. A higher value denotes higher change. The general change of the overall political opinion is defined as the Euclidean Distance between the overall opinion before and after the meeting. The measure is described in Section 1.5 and regression specifications are detailed in Section 1.4. 95% confidence intervals are included.

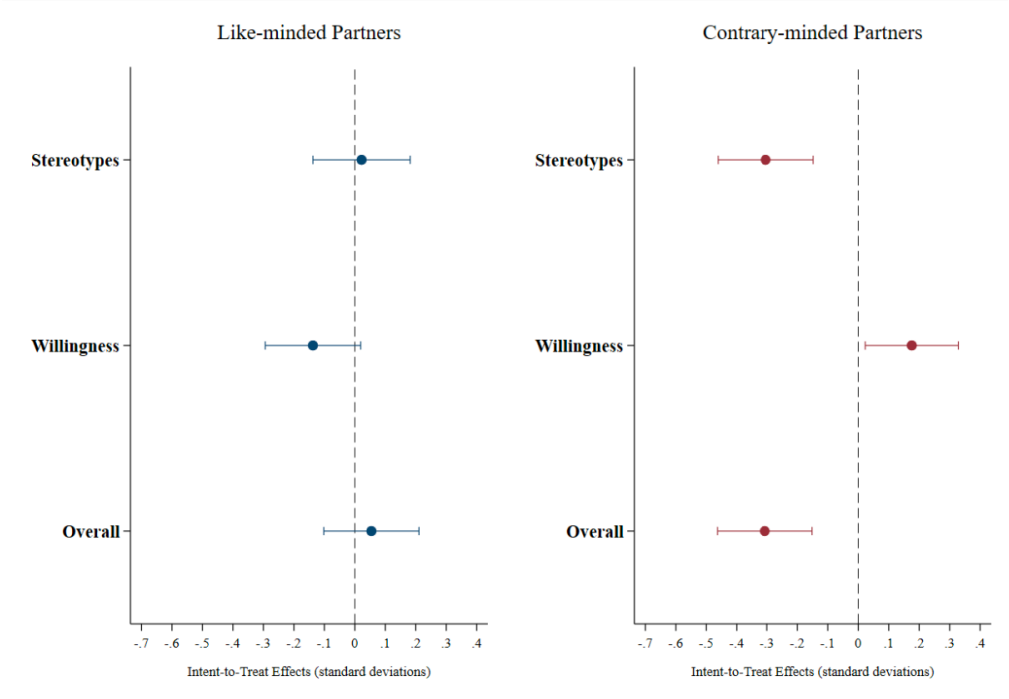


Figure 1.C.5. Effect of the Conversations on Affective Polarization (PDS)

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on affective polarization for the post double selection method (PDS). It plots the effects on (i) standardized overall stereotypes about a person with opposing political views, (ii) standardized willingness to engage in personal contact with a person that has opposing political views, and (iii) standardized overall affective polarization. The overall stereotype measure is defined as the first principal component of all four elicited stereotypes. Table 1.D.21 shows the loadings. Lower values denote lower stereotypes (and lower affective polarization). Lower willingness to engage in personal contact is associated with higher affective polarization. The overall affective polarization measure is defined as the first principal component of all four elicited stereotypes and the willingness to engage in personal contact. Table 1.D.29 shows the respective loadings. Lower values are associated with lower affective polarization. The measures are described in Section 1.6 and regression specifications are detailed in Section 1.4. 95% confidence intervals are included.

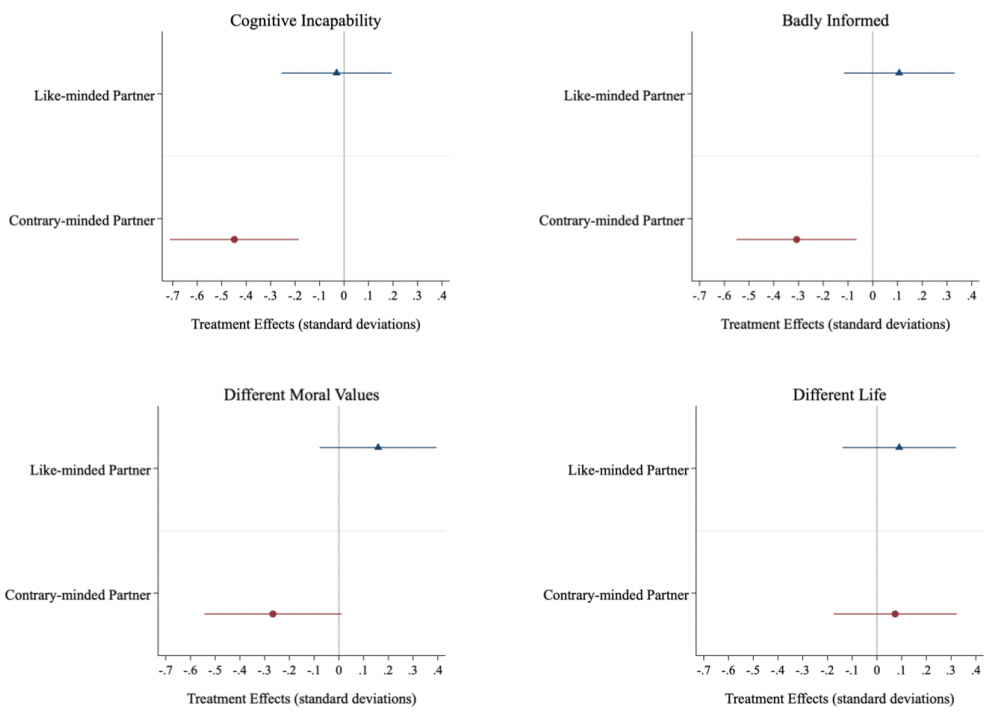


Figure 1.C.6. Effect on Stereotypes (Separate)

Notes: The figure shows the ITT effect of the like- and contrary-minded treatments on standardized stereotypes. Higher values denote higher stereotypes. The first panel shows the effect on the stereotype that contrary-minded individuals are cognitively less capable. The second panel plots the effect on the stereotype that contrary-minded individuals are poorly informed. The third and fourth panel show the effects on the stereotypes that contrary-minded individuals have different moral values and live completely different lifes, respectively. The measures are described in Section 1.6 and regression specifications are detailed in Section 1.4.

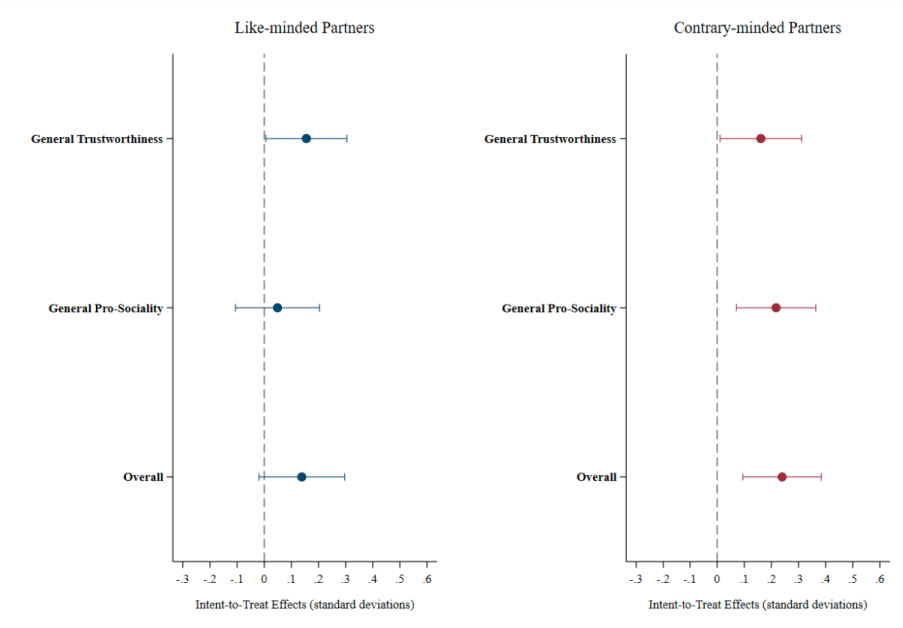


Figure 1.C.7. Effect of the Conversations on the Perception of Social Cohesion (PDS)

Notes: This figure shows the ITT effects of the like-minded treatments (left panel) and contrary-minded treatments (right panel) on standardized measures of perceptions of social cohesion for the post double selection method (PDS). It plots the impacts on (i) the perception that fellow citizens are generally trustworthy, (ii) the perception to what extent fellow citizens generally care about the well-being of other and (iii) the overall effect, defined as the first principal component of the two primer measures. Higher values denote higher perceptions. The outcome measures are described in Section 1.7 and regression specifications are detailed in Section 1.4. Error bars reflect 95% confidence intervals.

Appendix 1.D Additional Tables

Table 1.D.1. Political Registration Questions

Question	Abbreviation
Can Muslims and Non-Muslims coexist in Germany?	Coexistence
Did the public debate about sexual harassment and #metoo have any positive effects?	Pos. Effects of #metoo
Should meat be taxed higher in order to reduce its consumption?	Tax on Meat
Should German city centers become car-free?	Car-free City Centers
Should Germany implement stricter border controls?	Stricter Border Control
Are Germans worse off today than 10 years ago?	Germans worse off
Is Donald Trump good for the USA?	Trump: Good for USA

Notes: The table lists all seven political registration questions. The answers were elicited during registration and served as the basis for the matching with the partners. The answer scale was binary.

Table 1.D.2. Five Open Questions

Question / Statement
What do you do for a living?
You are a friend of....
What do you do in your free time?
How would you describe yourself?
What are your dislikes?

Notes: The table shows the five open questions elicited during registration for *Germany Talks*.

Table 1.D.3. Membership of Participants of *Germany Talks* to "Left" and "Right" Class

	Class 1: Left Ideology (kmeans)	Class 2: Right Ideology (kmeans)
Class 1: Left Ideology (LCA)	15,721	0
Class 2: Right Ideology (LCA)	377	2997

Notes: This table shows the number of participants of *Germany Talks* who belong to either the "left" or the "right" class, identified by LCA (rows) and k-means clustering (columns), respectively. The LCA is discussed in Section 1.3.

Table 1.D.4. Like-minded vs contrary-minded Partners

	Like-minded Partner (%)	Contrary-minded Partner(%)
Gender		
Female	38	21
Male	62	79
Age		
18 - 34	46	33
35 - 54	34	38
55 or older	21	29
Ideological Class		
Left Ideology	98	57
Right Ideology	2	43
Ideological Class: Overlap		
Same Ideological Class	97	26
Different Ideological Class	3	74

Notes: This table summarizes the characteristics of the partners in the like-minded LM (column 1) and the contrary-minded CM treatment condition (column 2). As most partners did not fill out the surveys, only age, gender and ideological (LCA) classes are available. Class membership is defined by the answers to the political registration questions. The last two rows indicate whether the two partners within one pair belong to the same class or not. The LCA is described in Section 1.3.

Table 1.D.5. Balance Checks

	Like-minded Partner	Contrary-minded Partner
Political Attitudes		
Border Control	0.137	-0.0270
	(0.131)	(0.138)
#metoo	-0.151	-0.212
	(0.121)	(0.145)
Meat Tax	0.00334	-0.0752
	(0.140)	(0.189)
Car free inner-cities	-0.174	-0.170
	(0.130)	(0.160)
Coexistence (Non-)Muslims	-0.0590	0.0679
	(0.110)	(0.149)
Germans worse off	0.0688	0.147
	(0.144)	(0.168)
Trump	-0.0204	0.149
	(0.103)	(0.122)
Same-sex marriage	-0.0505	0.0666
	(0.140)	(0.170)
Cooperation within EU	-0.0733	0.114
	(0.0886)	(0.120)
Income Tax	0.0764	-0.0690
	(0.160)	(0.181)
Trustworthiness Media	0.0547	-0.257
	(0.153)	(0.161)
Importance		
Importance: Border Control	0.0639	0.193
	(0.209)	(0.220)
Importance: #metoo	0.0827	-0.141
	(0.163)	(0.195)
Importance: Meat Tax	0.0190	0.0870
	(0.165)	(0.184)
Importance: Car free inner-cities	0.0349	0.0444
	(0.167)	(0.191)
Importance: Coexistence (Non-)Muslims	0.169	0.142
	(0.151)	(0.156)
Importance: Germans worse off	0.351*	0.163
	(0.207)	(0.203)
Importance: Trump	0.305	-0.0309
	(0.210)	(0.222)
Beliefs		
Number applications for asylum	-16025.4	-6738.0
	(32060.3)	(37974.5)
Share Muslims in Population	-0.0148	0.125
	(0.562)	(0.696)
Political Engagement		
Participation in citizens' initiative	0.0284	-0.0148
	(0.0272)	(0.0313)
Participation in demonstration	-0.0905*	-0.0102
	(0.0528)	(0.0500)
Work for party	0.0460	0.00946
	(0.0358)	(0.0448)
Work for union	0.0183	-0.00592
	(0.0215)	(0.0261)
None	0.00981	0.00447
	(0.0523)	(0.0573)
Not specified	-0.0119	0.0170
	(0.0158)	(0.0157)
Marital Status		
Single	0.00486	-0.0288
	(0.0419)	(0.0450)
Single, in relationship	-0.00394	0.0225
	(0.0417)	(0.0501)

Table 1.D.5. (continued)

	Like-minded Partner	Contrary-minded Partner
Life Partnership	-0.00686	-0.00538
	(0.0109)	(0.00768)
Married	-0.0614	-0.00108
	(0.0472)	(0.0536)
Married, living separately	0.0308	-0.00321
	(0.0215)	(0.0167)
Divorced	0.0261	0.0131
	(0.0216)	(0.0334)
Widowed	-0.00449	0.00783
	(0.0139)	(0.0146)
Not specified	0.0137	-0.00495
	(0.0160)	(0.0120)
Social Environment		
No one	0.0208*	-0.00855
	(0.0122)	(0.00569)
Almost no one	-0.0572	-0.0137
	(0.0348)	(0.0411)
Some	-0.0100	0.102*
	(0.0543)	(0.0607)
Approx. half	0.0635	-0.0442
	(0.0431)	(0.0535)
Many	-0.0326	-0.0422
	(0.0339)	(0.0389)
Almost everyone	0.00731	0.00310
	(0.00593)	(0.0143)
Religion		
None	-0.0521	-0.0122
	(0.0522)	(0.0563)
Christian	0.0313	0.0171
	(0.0515)	(0.0539)
Other	-0.00654	0.00755
	(0.0149)	(0.0174)
Not Specified	0.0274*	-0.0125
	(0.0154)	(0.0168)
Religiousness		
Never	-0.0602	-0.0485
	(0.0517)	(0.0600)
Less than several times per year	0.00396	0.0203
	(0.0550)	(0.0592)
Several times per year	0.0519	0.0144
	(0.0415)	(0.0410)
One to three times per month	0.0106	-0.00417
	(0.0217)	(0.0264)
Once per week	-0.0100	0.0238
	(0.0164)	(0.0151)
Several times per week	-0.00836	0.00920
	(0.0169)	(0.0120)
Not specified	0.0121	-0.0150
	(0.0102)	(0.0146)
Political spectrum left-right		
Far-left	-0.0349	0.00198
	(0.0212)	(0.0248)
Left	0.0226	-0.0419
	(0.0515)	(0.0476)
Centre-left	-0.00872	0.0395
	(0.0544)	(0.0569)
Centre	0.0627	-0.0103
	(0.0384)	(0.0466)
Centre-right	-0.0317	0.0224
	(0.0220)	(0.0352)
Right	-0.0000269	-0.00797

Table 1.D.5. (continued)

	Like-minded Partner	Contrary-minded Partner
	(0.00119)	(0.0185)
Far right	0.00128	0.00814
	(0.00228)	(0.0131)
Not specified	-0.0113	-0.0119
	(0.0121)	(0.0112)
Party		
CDU/CSU	-0.0273	0.00295
	(0.0208)	(0.0292)
SPD	0.0564*	-0.0197
	(0.0325)	(0.0356)
Bündnis/90 Die Grüne	-0.0145	-0.0413
	(0.0536)	(0.0561)
FDP	0.0247	0.00427
	(0.0234)	(0.0349)
Die Linke	-0.0653	-0.0206
	(0.0396)	(0.0402)
AfD	-0.000179	0.0518**
	(0.00153)	(0.0230)
Other party	0.0154	0.0162
	(0.0185)	(0.0287)
Don't Vote	0.000221	-0.000299
	(0.00632)	(0.00877)
Not specified	0.0107	0.00673
	(0.0215)	(0.0270)
F-Test	1.11	1.12
P-Value	0.28	0.27

Notes: The table reports the treatment coefficients of the balance checks if only the set of basic controls is conditioned on. Dependent variables are measures from the baseline survey: Baseline political views, subjective evaluation of importance of political topics, baseline beliefs about the share of muslims in Germany and number of asylum seekers in Germany, and baseline values of the additional set of controls. Each of these variables is regressed on the treatment dummy and the sets of basic controls. The respective dependent variable is listed in the first column. Column (1) reports the results for the like-minded and column (2) for the contrary-minded individuals. F-Tests are calculated by regressing the treatment on all those variables and the sets of basic controls. Robust standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.6. Attrition

	Like-minded Condition (LM)	Contrary-minded Condition (CM)
	(1)	(2)
Treat	-0.0162	-0.0228
	(0.0345)	(0.0357)
Constant	0.845**	0.640
	(0.365)	(0.393)
Basic Controls (no income)	Yes	Yes
Add. Controls (no marital st.)	Yes	Yes
Outcome Mean	0.49	0.49
Observations	1489	1412

Notes: Regression estimates, robust standard errors in parentheses. Dependent variable is a dummy variable equal to one if the participant filled out the baseline survey but did not complete the endline survey. It is equal to zero if only the baseline was completed. Column (1) shows the results for the like-minded treatment condition, column (2) for the contrary-minded treatment condition. Income and marital status were elicited in the endline survey and thus not conditioned onn. As the specification used here differs from the specification discussed in Section 1.4, results should be interpreted cautiously with respect to the existence of selective attrition. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$

Table 1.D.7. Selective Response Rate (Panel)

	Like-minded Condition (LM)	Contrary-minded Condition (CM)
	(1)	(2)
Treat	0.0669***	0.0715***
	(0.0126)	(0.0155)
Constant	-0.0449	0.494***
	(0.0685)	(0.152)
Basic Controls (in parts)	Yes	Yes
Outcome Mean	0.189	0.215
Observations	4032	3391

Notes: Regression estimates. Dependent variable is a dummy variable equal to one if the participant filled out both surveys and equal to zero if no survey was completed. Column (1) shows the results for like-minded treatment condition, column (2) for the contrary-minded treatment condition. Treat is a dummy that equals to one if the first-accepter and the partner accepted, and zero otherwise. Income, education, migration background (basic controls) and all additional controls were elicited in the endline survey and thus not conditioned on. As the specification used here differs very much from the specification discussed in Section 1.4, results should be interpreted cautiously with respect to the existence of selective response. Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 1.D.8. Political Distance Dependent Selection

	All Participants	Panel
	(1)	(2)
Contrary-minded	-0.00553	0.0157
	(0.00721)	(0.0187)
Constant	0.446***	0.633***
	(0.00488)	(0.0131)
R^2	0.0000307	0.000267
Observations	19135	2646

Notes: The table reports OLS estimates. The dependent variable is a dummy equal to one if a person accepted first and zero if she did not accepted or accepted second. *Contrary-minded* is 1 if the participant was assigned to a contrary-minded partner. The first column contains all available observations while in column (2) the sample is restricted to people who answered both surveys. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.9. Change towards Extreme Views: Absolute (Dis-)Agreement

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.283**	0.286**	0.281***	-0.0645	-0.0225	-0.00911
	(0.123)	(0.127)	(0.0810)	(0.140)	(0.151)	(0.0827)
Constant	0.615	0.927		-2.226**	-1.309	
	(0.575)	(1.133)		(0.924)	(1.777)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	721	721	721	695	695	695
R^2	0.386	0.447		0.521	0.582	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views in terms of absolute (dis-)agreement. Positive coefficients mean a change towards more extreme views. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): Two combinations of the political registration questions, various NUTS FE. The outcome measure is described in Section 1.5 and regression specifications are detailed in Section 1.4.*$p < 0.10$, **$p < 0.05$, and ***$p < 0.01$

Table 1.D.10. Change towards Extreme Views (Relative to Population)

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.232*	0.279**	0.199**	-0.156	-0.151	-0.112
	(0.130)	(0.129)	(0.0834)	(0.131)	(0.141)	(0.0797)
Constant	1.505**	1.256		-2.404**	-4.168**	
	(0.738)	(1.255)		(0.959)	(1.664)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	721	721	721	695	695	695
R^2	0.381	0.448		0.540	0.585	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views relative to the population. Positive coefficients indicate a change towards more extreme views. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): One combination of the political registration questions, various NUTS FE. The outcome measure is described in Section 1.5 and regression specifications are detailed in Section 1.4.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.11. Effect on Ideological Polarization (Extreme Views): Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.258**	0.276**	0.243***	-0.122	-0.00411	0.00523
	(0.111)	(0.114)	(0.0751)	(0.176)	(0.200)	(0.0912)
Constant	-0.632	-0.268		-2.321**	-0.556	
	(0.573)	(1.166)		(0.909)	(2.038)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	876	876	876	540	540	540
R^2	0.309	0.368		0.596	0.694	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views in terms of absolute (dis-)agreement. Treatment conditions are defined by using overlap of ideological classes (see Section 1.3). Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: . The outcome measure is described in Section 1.5 and regression specifications are detailed in Section 1.4.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.12. Effect on Ideological Polarization (Non-average Views): Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.226**	0.286**	0.187**	-0.221	-0.102	-0.171*
	(0.108)	(0.111)	(0.0738)	(0.174)	(0.184)	(0.0916)
Constant	-0.220	-0.763		-2.750***	-2.789	
	(0.602)	(1.158)		(0.995)	(1.943)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	876	876	876	540	540	540
R^2	0.322	0.385		0.651	0.734	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized ideological polarization towards non-average views. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Treatment conditions are defined by using overlap of ideological classes (see Section 1.3) Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: . The outcome measure is described in Section 1.5 and regression specifications are detailed in Section 1.4.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.13. Alt. Treatment Conditions: Comparison of Different Cut-Offs

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	(1) Standard LM	(2) Weak LM	(3) Strict LM	(4) Standard CM	(5) Strict CM	(6) Weak CM
Abs. (Dis-)Agreement	0.286**	0.344**	0.270**	-0.0225	0.0658	-0.127
	(0.127)	(0.159)	(0.106)	(0.151)	(0.118)	(0.272)
Rel. to Population	0.279**	0.323*	0.245**	-0.151	-0.0331	-0.256
	(0.129)	(0.179)	(0.110)	(0.141)	(0.112)	(0.230)
Stereotypes	0.0873	0.185	0.0554	-0.379***	-0.237**	-0.552**
	(0.122)	(0.165)	(0.0960)	(0.132)	(0.0966)	(0.230)
Willingness Contact	-0.0993	-0.0994	-0.114	0.146	0.0208	0.160
	(0.115)	(0.147)	(0.0906)	(0.133)	(0.101)	(0.212)
Trustworthiness	0.114	0.0366	0.159*	0.274**	0.253***	0.400*
	(0.122)	(0.168)	(0.0897)	(0.114)	(0.0872)	(0.204)
Pro-Sociality	0.0438	0.0412	0.0629	0.245**	0.176*	0.208
	(0.125)	(0.175)	(0.0939)	(0.117)	(0.0943)	(0.226)

Notes: Regression estimates, robust standard errors in parentheses. Treatment coefficients are reported. The dependent variable are standardized change towards extreme views (rows 1 and 2), stereotypes and willingness to engage in personal contact (rows 3 and 4), and the beliefs of trustworthiness and pro-sociality (rows 5 and 6). Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1) and (4) show the results for the standard split into the like- and contrary-minded condition. Columns (2) and (5) report the results if first-accepters are assigned to the like-minded (contrary-minded) condition only if they answered 2 (3) or less (more) of the political registration questions differently. Columns (3) and (6) report the results if first-accepters are assigned to the like-minded (contrary-minded) condition only if they answered 4 (2) or less (more) of the political registration questions differently. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.14. Change towards Extreme Views: Abs. (Dis-)Agreement - Manhattan Dist.

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.262**	0.264**	0.237***	-0.0908	-0.0392	-0.0663
	(0.121)	(0.129)	(0.0819)	(0.137)	(0.145)	(0.0823)
Constant	0.442	1.129		-2.792***	-2.558	
	(0.705)	(1.217)		(0.886)	(1.691)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	721	721	721	695	695	695
R^2	0.376	0.437		0.532	0.599	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views in terms of absolute (dis-)agreement, measured with the Manhattan Distance. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): Two combinations of the political registration questions, various NUTS FE.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.15. Change towards Extreme Views: Abs.(Dis-)Agreement - Mahalanobis Dist.

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.181	0.213*	0.232***	-0.0402	-0.0278	-0.0455
	(0.117)	(0.120)	(0.0799)	(0.152)	(0.163)	(0.0846)
Constant	0.137	-0.533		-2.455**	-3.181*	
	(0.592)	(1.163)		(1.110)	(1.902)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	721	721	721	695	695	695
R^2	0.412	0.478		0.492	0.562	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views in terms of absolute (dis-)agreement, measured with the Mahalanobis Distance. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): Two combinations of the political registration questions, various NUTS FE.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.16. Change towards Extreme Views (Rel. to Population) - Manhattan Dist.

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.201	0.233*	0.173**	-0.173	-0.152	-0.160**
	(0.126)	(0.128)	(0.0830)	(0.126)	(0.132)	(0.0792)
Constant	1.930*	1.967		-2.416***	-3.793**	
	(1.046)	(1.332)		(0.860)	(1.600)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	721	721	721	695	695	695
R^2	0.381	0.443		0.565	0.612	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views relative to the population, measured with the Manhattan Distance. Positive coefficients mean adjustment away from the average opinion, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): One combination of the political registration questions, various NUTS FE.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.17. Change towards Extreme Views (Rel. to Population) - Mahalanobis Dist.

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.145	0.200	0.148*	-0.131	-0.139	-0.0989
	(0.131)	(0.134)	(0.0839)	(0.128)	(0.138)	(0.0809)
Constant	1.255	0.516		-2.560***	-4.447**	
	(0.862)	(1.216)		(0.964)	(1.879)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	721	721	721	695	695	695
R^2	0.382	0.449		0.567	0.613	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized change towards more extreme views relative to the population, measured with the Mahalanobis Distance. Positive coefficients mean adjustment away from the average opinion, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): One combination of the political registration questions, various NUTS FE.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.18. Effect on Ideological Polarization (Reweighted)

	Like-minded				Contrary-minded			
	Absolute		Relative		Absolute		Relative	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.286**	0.282**	0.279**	0.262*	-0.0225	-0.0614	-0.151	-0.234
	(0.127)	(0.134)	(0.129)	(0.137)	(0.151)	(0.166)	(0.141)	(0.156)
Constant	0.927	1.271	1.256	1.408	-1.309	-0.900	-4.168**	-4.919**
	(1.133)	(1.090)	(1.255)	(1.320)	(1.777)	(1.916)	(1.664)	(2.175)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reweighted	No	Yes	No	Yes	No	Yes	No	Yes
Observations	721	721	721	721	695	695	695	695
R^2	0.447	0.568	0.448	0.571	0.582	0.592	0.585	0.591

Notes: The table reports ITT effects of in-person conversations on the two standardized ideological polarization measures, change towards extreme views in terms of absolute (dis-)agreement (columns 1, 2, 5, 6) and relative to the population (columns 3, 4, 7, 8). Columns (1), (3), (5) and (7) show the estimates using equal weights. These columns are the same as columns (2) and (5) in Table 1.D.9 and Table 1.D.10, respectively. Columns (2) and (4) reweight the like-minded subsample to match the contrary-minded subsample on the following covariates: mean age, share of males, females and non-binary, party shares, and self-reported left-right classification. Analogously, Columns (6) and (8) reweight the contrary-minded subsample to match the like-minded subsample on the these covariates. This analysis is discussed in Section 1.4. Robust standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.19. Effect on Attitudes: General Adjustment

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.303***	0.294**	0.216***	0.0998	0.0700	0.167**
	(0.115)	(0.122)	(0.0818)	(0.143)	(0.138)	(0.0790)
Constant	0.664	0.373		-0.738	-4.155**	
	(0.791)	(1.379)		(1.167)	(2.037)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	721	721	721	695	695	695
R^2	0.405	0.459		0.535	0.615	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized general change. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Two combinations of the political registration questions, various NUTS FE. Column (6): One combination of the political registration questions, various NUTS FE, one social environment dummy. The outcome measure is described in Section 1.5 and regression specifications are detailed in Section 1.4.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.20. Effect on Stereotypes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0847	0.0873	0.0303	-0.292**	-0.379***	-0.305***
	(0.117)	(0.122)	(0.0814)	(0.120)	(0.132)	(0.0798)
Constant	-2.542**	-2.519*		-2.496**	-2.421	
	(1.196)	(1.412)		(0.982)	(1.489)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	747	747	747	720	720	720
R^2	0.388	0.470		0.561	0.618	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized stereotypes about contrary-minded. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): Two combinations of the political registration questions, two NUTS FE..* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.21. PCA: Loadings Stereotypes on Principal Component

Stereotype	Loadings
Different Way of Life	0.36
Different Moral Values	0.33
Low Cognitive Abilities	0.61
Poorly Informed	0.62

*Notes:*The table presents the loadings of the principal component analysis of all four stereotypes on the first principal component. The first component is the linear combination of the four stereotypes with the respective loadings as weights.

Table 1.D.22. Effect on Stereotypes: Different Way of Life

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.150	0.0903	0.113	0.0853	0.0738	-0.0552
	(0.107)	(0.116)	(0.0752)	(0.125)	(0.127)	(0.0799)
Constant	-1.927***	-1.788*		-1.301	-1.012	
	(0.557)	(1.039)		(0.838)	(1.609)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes	X	No	Yes	X
Observations	755	755	755	725	725	725
R^2	0.420	0.479		0.536	0.616	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized belief that contrary-minded lead a different way of life. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): One combination of the political registration questions, various NUTS FE, one education dummy. Column (6): Two combinations of the political registration questions, one NUTS FE, one social environment dummy.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.23. Effect on Stereotypes: Different Moral Values

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.142	0.159	0.0897	-0.214	-0.267*	-0.234***
	(0.111)	(0.120)	(0.0765)	(0.130)	(0.141)	(0.0797)
Constant	-0.704	-0.370		-1.718*	-0.796	
	(0.903)	(1.215)		(0.969)	(1.741)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	753	753	753	725	725	725
R^2	0.368	0.439		0.503	0.570	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized belief that contrary-minded individuals have different moral values. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE. Column (6): One combination of the political registration questions, one NUTS FE.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.24. Effect on Stereotypes: Low Cognitive Abilities

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-0.0414	-0.0305	-0.0595	-0.366***	-0.448***	-0.341***
	(0.110)	(0.115)	(0.0765)	(0.124)	(0.134)	(0.0809)
Constant	-1.819*	-2.095*		-1.594	-1.327	
	(1.039)	(1.202)		(1.000)	(1.557)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes	X	No	Yes	
Observations	753	753	753	725	725	725
R^2	0.372	0.439		0.529	0.586	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized belief that contrary-minded individuals have low cognitive abilities. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE, one education dummy.Column (6):Two combinations of the political registration questions.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.25. Effect on Stereotypes: Poorly Informed

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0733	0.107	0.0246	-0.228*	-0.308**	-0.144*
	(0.110)	(0.114)	(0.0771)	(0.116)	(0.123)	(0.0784)
Constant	-2.410**	-2.355		-1.987**	-2.777**	
	(1.213)	(1.454)		(0.967)	(1.345)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes	X	No	Yes	X
Observations	753	753	753	726	726	726
R^2	0.380	0.464		0.562	0.626	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized belief that contrary-minded individuals are poorly informed. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): Various NUTS FE, one party dummy. Column (6): One combination of the political registration questions, two NUTS FE, one income dummy.* * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.26. Effect on Stereotypes: Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0563	0.0426	0.00151	-0.341**	-0.388**	-0.328***
	(0.0938)	(0.0970)	(0.0709)	(0.162)	(0.177)	(0.0930)
Constant	-3.424***	-3.943***		-2.194**	-1.107	
	(0.585)	(0.955)		(1.091)	(1.714)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes	X	No	Yes	X
Observations	910	910	910	557	557	557
R^2	0.383	0.450		0.643	0.716	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized overall stereotype measure. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: .* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.27. Effect on Affective Polarization (Reweighted)

	Like-minded				Contrary-minded			
	Stereotypes		Willingness		Stereotypes		Willingness	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.0873	0.120	-0.0993	-0.0929	-0.379***	-0.469***	0.146	0.240*
	(0.122)	(0.122)	(0.115)	(0.117)	(0.132)	(0.141)	(0.133)	(0.137)
Constant	-2.519*	-2.572*	-0.563	-0.822	-2.421	-2.876	0.211	-0.0932
	(1.412)	(1.385)	(1.104)	(1.072)	(1.489)	(1.769)	(1.482)	(1.802)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reweighted	No	Yes	No	Yes	No	Yes	No	Yes
Observations	747	747	755	755	720	720	727	727
R^2	0.470	0.567	0.501	0.609	0.618	0.629	0.582	0.611

Notes: The table reports ITT effects of in-person conversations on the two standardized affective polarization measures, overall stereotypes (columns 1, 2, 5, 6) and willingness to engage in personal contact (columns 3, 4, 7, 8). Columns (1), (3), (5) and (7) show the estimates using equal weights. These columns are the same as columns (2) and (5) in Table 1.D.20 and Table 1.5, respectively. Columns (2) and (4) reweight the like-minded subsample to match the contrary-minded subsample on the following covariates: mean age, share of males, females and non-binary, party shares, and self-reported left-right classification. Analogously, Columns (6) and (8) reweight the contrary-minded subsample to match the like-minded subsample on the these covariates. This analysis is discussed in Section 1.4. Robust standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.28. Willingness to Engage in Personal Contact: Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-0.150	-0.166*	-0.129*	0.219	0.235	0.232**
	(0.0918)	(0.0952)	(0.0682)	(0.152)	(0.165)	(0.0927)
Constant	-0.596	-0.944		-0.0648	-1.821	
	(0.553)	(0.852)		(1.009)	(1.673)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	
Observations	918	918	918	564	564	564
R^2	0.336	0.418		0.649	0.696	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized overall stereotype measure. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: .* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.29. PCA: Loadings Stereotypes and Willingness to Engage in Personal Contact on First Principal Component

Stereotype	Loadings
Different Way of Life	0.34
Different Moral Values	0.32
Low Cognitive Abilities	0.54
Poorly Informed	0.55
Willingness to Engage in Personal Contact	-0.43

Notes: The table presents the loadings of the principal component analysis of all four stereotypes and willingness to engage in personal contact on the first principal component which denotes our measure for overall affective polarization. The first component is the linear combination of the four stereotypes and willingness with the respective loadings as weights. The loadings are consistent with an interpretation of the component as an overall affective polarization measure as the signs of the loadings are positive for stereotypes and negative for willingness.

Table 1.D.30. Effect on Perception of General Trustworthiness

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0963	0.114	0.163**	0.229**	0.274**	0.155**
	(0.114)	(0.122)	(0.0768)	(0.109)	(0.114)	(0.0761)
Constant	-1.259	-2.196		-0.502	-0.948	
	(1.259)	(1.413)		(0.889)	(1.852)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	757	757	757	726	726	726
R^2	0.356	0.430		0.655	0.698	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized perception of general trustworthiness. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): One combination of the political registration questions, various NUTS FE, one income dummy. Column (6): Two combinations of the political registration questions, various NUTS FE, one income dummy, two political party dummies.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.31. Effect on Perception of General Pro-Sociality

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0211	0.0438	0.0585	0.255**	0.245**	0.217***
	(0.114)	(0.125)	(0.0786)	(0.109)	(0.117)	(0.0746)
Constant	-1.078	-0.107		0.960	1.566	
	(1.248)	(1.037)		(0.815)	(1.536)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	759	759	759	727	727	727
R^2	0.384	0.456		0.595	0.657	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized perception of general pro-sociality. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: Column (3): One combination of the political registration questions, two NUTS FE, one education dummy. Column (6): Various combinations of the political registration questions, two NUTS FE, one political party dummy.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.32. Effect on Perception of General Trustworthiness: Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.101	0.131	0.132**	0.309**	0.252	0.201**
	(0.0858)	(0.0887)	(0.0644)	(0.151)	(0.166)	(0.0946)
Constant	-1.283*	-0.701		-0.494	-0.722	
	(0.722)	(1.244)		(1.143)	(2.287)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	921	921	921	562	562	562
R^2	0.321	0.376		0.690	0.738	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is standardized perception of trustworthiness. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: .* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.33. Effect on Perception of General Pro-Sociality: Ideological Classes

	Like-minded Partner (LM)			Contrary-minded Partner (CM)		
	OLS	OLS	PDS	OLS	OLS	PDS
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.0612	0.0753	0.0839	0.310**	0.273*	0.218**
	(0.0897)	(0.0934)	(0.0659)	(0.146)	(0.161)	(0.0895)
Constant	-2.020***	-1.417		1.516*	2.828	
	(0.554)	(1.106)		(0.858)	(1.858)	
Basic Controls	Yes	Yes	X	Yes	Yes	X
Additional Controls	No	Yes		No	Yes	X
Observations	923	923	923	563	563	563
R^2	0.356	0.428		0.631	0.692	

Notes: Regression estimates, robust standard errors in parentheses. The dependent variable is the standardized belief about general pro-sociality. Positive coefficients mean adjustment away from the center towards the boundary, negative coefficients the opposite. Columns (1) - (3) report the results for those with like-minded partners (LM), columns (4) - (6) for those with contrary-minded partners (CM). Columns (1), (2), (4) and (5) present OLS and columns (3) and (6) PDS regressions. Basic controls include dummies for age intervals, gender, NUTS regions, combinations of seven political registration questions, education, income and migration background. Additional controls consist of dummies for political parties, political self-classification, political engagement, religion, religiousness, marital status, and number of politically contrary-minded people in social environment. Variables selected by the PDS procedure (denoted by "X") are: .* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.34. Effect on Perception of Social Cohesion (Reweighted)

	Like-minded				Contrary-minded			
	Trustworthiness		Pro-Sociality		Trustworthiness		Pro-Sociality	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.114	0.0731	0.0438	-0.0234	0.274**	0.186*	0.245**	0.166
	(0.122)	(0.132)	(0.125)	(0.125)	(0.114)	(0.108)	(0.117)	(0.124)
Constant	-2.196	-2.748*	-0.107	-0.163	-0.948	-1.496	1.566	0.367
	(1.413)	(1.404)	(1.037)	(0.979)	(1.852)	(2.378)	(1.536)	(2.300)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reweighted	No	Yes	No	Yes	No	Yes	No	Yes
Observations	757	757	759	759	726	726	727	727
R^2	0.430	0.493	0.456	0.553	0.698	0.676	0.657	0.643

Notes: The table reports ITT effects of in-person conversations on standardized perceptions of general trustworthiness (columns 1, 2, 5, 6) and general pro-sociality (columns 3, 4, 7, 8). Columns (1), (3), (5) and (7) show the estimates using equal weights. These columns are the same as columns (2) and (5) in Table 1.D.30 and Table 1.D.31, respectively. Columns (2) and (4) reweight the like-minded subsample to match the contrary-minded subsample on the following covariates: mean age, share of males, females and non-binary, party shares, and self-reported left-right classification. Analogously, Columns (6) and (8) reweight the contrary-minded subsample to match the like-minded subsample on the these covariates. This analysis is discussed in Section 1.4. Robust standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.35. Disappointment: Comparison of Time Trends

	Affective Polarization				Social Coehsion			
	Stereotypes		Willingness		Trustworthines		Pro-Sociality	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Time × CM	0.0298	0.0112	-0.0641	-0.0316	-0.0139	-0.0336	-0.231*	-0.207
	(0.0999)	(0.121)	(0.120)	(0.145)	(0.0916)	(0.110)	(0.120)	(0.146)
CM	0.000499	-0.0697	0.228	-0.0191	-0.257**	-0.0563	-0.343***	0.00309
	(0.129)	(0.182)	(0.140)	(0.209)	(0.112)	(0.152)	(0.125)	(0.180)
Time	0.190**	0.188**	-0.198**	-0.212**	0.271***	0.290***	0.166**	0.173*
	(0.0734)	(0.0894)	(0.0879)	(0.107)	(0.0679)	(0.0819)	(0.0841)	(0.102)
Constant	-0.204**	-1.257	3.448***	4.039**	4.089***	-0.217	3.460***	2.810*
	(0.0909)	(1.430)	(0.100)	(1.585)	(0.0795)	(1.409)	(0.0893)	(1.436)
Basic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Additional Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	1090	1075	1098	1083	1098	1083	1100	1085

Notes: The table tests for different time trends between the control groups. It shows regression results of the non-standardized outcome variables on the dummy *time*, the dummy *CM* and their interaction. *CM* denotes whether a person was matched to a like- or a contrary-minded partner. Standard errors are clustered at participant level. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

References

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review* 110 (3): 629–76. [1, 3, 5, 23, 24]

Allport, Gordon Willard. 1954. "The nature of prejudice." [5, 24]

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216–9221. [5, 19]

Baysan, Ceren. 2021. "Persistent Polarizing Effects of Persuasion: Experimental Evidence from Turkey." Working paper. Working paper. [21]

Beam, Michael A, Myiah J Hutchens, and Jay D Hmielowski. 2018. "Facebook news and (de) polarization: reinforcing spirals in the 2016 US election." *Information, Communication & Society* 21 (7): 940–958. [5]

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies* 81 (2): 608–650. [17]

Bishop, Bill. 2009. *The big sort: Why the clustering of like-minded America is tearing us apart*. Houghton Mifflin Harcourt. [4]

Boisjoly, Johanne, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles. 2006. "Empathy or antipathy? The impact of diversity." *American Economic Review* 96 (5): 1890–1905. [6]

Boxell, Levi. 2020. "Demographic change and political polarization in the United States." *Economics Letters* 192: 109187. [23, 24]

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2020. "Cross-country trends in affective polarization." Working paper. National Bureau of Economic Research. [1]

Broockman, David, and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352 (6282): 220–224. [2, 4, 27]

Brown, Jacob R, and Ryan D Enos. 2021. "The measurement of partisan sorting for 180 million voters." *Nature Human Behaviour*, 1–11. [4]

Burns, Justine, Lucia Corno, and Eliana La Ferrara. 2015. "Interaction, prejudice and performance. Evidence from South Africa." Working paper. [6]

Carrell, Scott E, Mark Hoekstra, and James E West. 2015. "The impact of intergroup contact on racial attitudes and revealed preferences." Working paper. National Bureau of Economic Research. [6]

Chen, M Keith, and Ryne Rohla. 2018. "The effect of partisanship and political advertising on close family ties." *Science* 360 (6392): 1020–1024. [22]

Di Tella, Rafael, Ramiro H Gálvez, and Ernesto Schargrodsky. 2021. "Does social media cause polarization? evidence from access to twitter echo chambers during the 2019 argentine presidential debate." Working paper. National Bureau of Economic Research. [5]

Dinesen, Peter Thisted, Merlin Schaeffer, and Kim Mannemar Sønderskov. 2020. "Ethnic diversity and social trust: A narrative and meta-analytical review." *Annual Review of Political Science* 23: 441–465. [28]

Eady, Gregory, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A Tucker. 2019. "How many people live in political bubbles on social media? Evidence from linked survey and Twitter data." *Sage Open* 9 (1): 2158244019832705. [5]

Finseraas, Henning, and Andreas Kotsadam. 2017. "Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment." *European Journal of Political Research* 56 (3): 703–722. [6]

Fishkin, J, A Siu, L Diamond, and N Bradburn. No date. "Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on America in One Room." *APSA Preprint*, (). [24]

Fishkin, James, Alice Siu, Larry Diamond, and Norman Bradburn. 2021. "Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on "America in One Room"." *American Political Science Review*, 1–18. [5]

Flaxman, Seth, Sharad Goel, and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80 (S1): 298–320. [5]

Freitag, Julian, Anna Kerkhof, and Johannes Münster. 2021. "Selective sharing of news items and the political position of news outlets." *Information Economics and Policy*, 100926. [31]

Gentzkow, Matthew. 2016. "Polarization in 2016." Working paper. Stanford University. [1]

Gentzkow, Matthew, and Jesse M Shapiro. 2011. "Ideological segregation online and offline." *Quarterly Journal of Economics* 126 (4): 1799–1839. [5]

Gerber, Alan S, and Donald P Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American political science review* 94 (3): 653–663. [2]

Green, Donald P, Alan S Gerber, and David W Nickerson. 2003. "Getting out the vote in local elections: Results from six door-to-door canvassing experiments." *Journal of Politics* 65 (4): 1083–1096. [2]

Gutmann, Amy, and Dennis F Thompson. 2009. *Why deliberative democracy?* Princeton University Press. [5]

Habermas, Jürgen. 1984. *The Theory of Communicative Action*. Beacon Press. [5]

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46. [19]

Halberstam, Yosh, and Brian Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics* 143: 73–88. [5]

Helbling, Marc, and Sebastian Jungkunz. 2020. "Social divides in the age of globalization." *West European Politics* 43 (6): 1187–1210. [6]

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual Review of Political Science* 22: 129–146. [1, 27]

Iyengar, Shanto, and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American Journal of Political Science* 59 (3): 690–707. [1, 6]

Kalla, Joshua L, and David E Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments." *American Political Science Review* 114 (2): 410–425. [2, 5, 24]

Levendusky, Matthew S. 2018. "Americans, not partisans: Can priming American national identity reduce affective polarization?" *Journal of Politics* 80 (1): 59–70. [5]

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review* 111 (3): 831–70. [5]

Lowe, Matt. 2021. "Types of contact: A field experiment on collaborative and adversarial caste integration." *American Economic Review* 111 (6): 1807–44. [6, 26, 28]

Martin, Gregory J, and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107 (9): 2565–99. [5]

Mason, Lilliana. 2015. ""I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization." *American Journal of Political Science* 59 (1): 128–145. [3]

Orr, Lilla V, and Gregory A Huber. 2020. "The policy basis of measured partisan animosity in the United States." *American Journal of Political Science* 64 (3): 569–586. [4]

Paluck, Elizabeth Levy. 2016. "How to overcome prejudice." *Science* 352 (6282): 147–147. [6]

Paluck, Elizabeth Levy, Seth A Green, and Donald P Green. 2019. "The contact hypothesis re-evaluated." *Behavioural Public Policy* 3 (2): 129–158. [2, 4, 6, 26]

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK. [5]

Peterson, Erik, Sharad Goel, and Shanto Iyengar. 2021. "Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign." *Political Science Research and Methods* 9 (2): 242–258. [1, 5]

Pettigrew, Thomas F, and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90 (5): 751. [2, 6]

PEW, Pew Research Center. 2014. *Political polarization in the American public*. Pew Research Center Washington, DC. [1]

PEW, Pew Research Center. 2018. *News Media and Political Attitudes in Germany*. Pew Research Center Washington, DC. [7, 31]

Pons, Vincent. 2018. "Will a five-minute discussion change your mind? A countrywide experiment on voter choice in France." *American Economic Review* 108 (6): 1322–63. [2]

Prior, Markus. 2013. "Media and political polarization." *Annual Review of Political Science* 16: 101–127. [5]

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review* 109 (3): 774–809. [6, 28]

Scacco, Alexandra, and Shana S Warren. 2018. "Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria." *American Political Science Review* 112 (3): 654–677. [6]

Schindler, David, and Mark Westcott. 2021. "Shocking racial attitudes: black GIs in Europe." *Review of Economic Studies* 88 (1): 489–520. [6]

Schkade, David, Cass R Sunstein, and Reid Hastie. 2007. "What happened on deliberation day." *Cal. L. Rev.* 95: 915. [5]

Simonsson, Otto, and Joseph Marks. 2020. "Love Thy (Partisan) Neighbor: Brief Befriending Meditation Reduces Affective Polarization." *Available at SSRN 3674051*, [5]

Sunstein, Cass R. 2009. *Going to extremes: How like minds unite and divide*. Oxford University Press. [1, 19]

Sunstein, Cass R. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press. [5]

Voelkel, Jan G, James Chu, Michael Stagnaro, Joe Mernyk, Chrystal Redekopp, Sophia Pink, James Druckman, David Rand, and Robb Willer. 2021. "Interventions Reducing Affective Polarization Do Not Improve Anti-Democratic Attitudes." Working paper. [5]

Warner, Benjamin R, Haley Kranstuber Horstman, and Cassandra C Kearney. 2020. "Reducing political polarization through narrative writing." *Journal of Applied Communication Research* 48 (4): 459–477. [5]

Wojcieszak, Magdalena. 2011. "Deliberation and attitude polarization." *Journal of Communication* 61 (4): 596–617. [19]

Chapter 2

Moral Luck: Mechanisms, Robustness and Prevalence

Joint with Armin Falk and David Huffman

2.1 Introduction

In many types of decisions, individuals can influence the probabilities of good or bad outcomes by their actions, but there is still a role for chance in determining what ultimately happens. For example, driving under the influence of alcohol may increase the probability of subsequently hitting and killing a pedestrian if a pedestrian crosses the street, but the presence of a pedestrian depends on chance. Likewise, an employee can take self-interested actions that expose the employer to increased risk of a loss, but chance will ultimately determine whether the loss occurs. More generally, in meritocratic societies, individuals can have a strong work ethic and exert high effort, but due to bad luck still end up being unsuccessful. In all of these cases, realized random outcomes do not contain any additional information about the intentions or effort of the actor beyond the observed actions. If punishments and rewards do vary with such outcomes, however, this violates a property of optimal incentives, sometimes called the "informativeness" principle (Holmström, 1979; Bolton and Dewatripont, 2004). Such violations would have profound implications for the functioning of legal systems, employment relationships, democracies, and meritocratic societies, by undermining the motivating and deterrent value of rewards and punishments.

It has been posited since ancient times that there is, in fact, a human tendency to reward or punish actors for outcomes that are beyond their control (Aristotle, 1984), a phenomenon sometimes called "moral luck," but the prevalence and robustness of this phenomenon are still not fully understood, and a key open question continues to be debated, which is whether moral luck is a preference or a bias. Understanding whether moral luck is a preference or a bias is important, because if

it is a preference, having punishments vary with outcomes satisfies some notion of what is appropriate or desired, which might offset the costs of providing suboptimal incentives. If the phenomenon is a bias, however, then this raises important questions about the desirability of how punishments and rewards are determined in many areas of society. It also points to a possible value of interventions to de-bias decisions. An early contributor to the debate on mechanisms was Adam Smith, who proposed that sentiments or emotions aroused by outcomes can affect perceptions of the actor, even though they contain no information about intentions, and thereby distort attributions of merit or demerit (Smith, 1790). Moral luck has been viewed as an ethical puzzle, as it seems to violate an appealing moral principle known as the "control principle," that requires that people only be held responsible for factors under their control (Kant, 1784). More modern philosophers, however, have continued to debate whether moral luck could be due to some coherent moral preference or principle that is an alternative to the control principle (e.g., Nagel, 1979; Williams, 1981; for a survey see Dana, 2021), and the question remains a current one for legal scholars (e.g., Enoch and Marmor, 2007).

This paper provides new evidence on the existence and prevalence of moral luck, and provides evidence on the question of mechanisms, indicating that it is at least partly a bias. As a first step we show evidence of moral luck in punishment behavior. Second, we show that random outcomes influence various judgments about the character of the actor, as well as incentivized beliefs about one aspect of the preferences of the actor, despite the random outcomes containing zero information. These biased judgements and beliefs are in turn correlated with punishment behavior. Third, to complete the causal chain, we exogenously vary whether punishers are provided with information about the actor's character, and show that this significantly reduces the influence of outcomes on punishment. This indicates that impact of random outcomes on beliefs about the actor is a mechanism underlying the variation of punishment with outcomes. We check robustness of the bias to an intervention that encourages deliberative rather than spontaneous decision making, and find that an influence of outcomes on beliefs and punishment remains, although the effects are somewhat smaller, indicating that the bias is relatively deep-seated and hard to remove. Interestingly, we also find that actors tend to internalize moral luck, in terms blaming themselves differently depending on random outcomes, which indicates that the phenomenon can emerge even when actions and outcomes are unobserved.

This study uses an approach that addresses some important methodological challenges to studying moral luck. With naturally occurring observational data on punishments and rewards, e.g., sentencing decisions from court cases, a problem is the difficulty of observing the roles of chance versus actions as perceived by those deciding on punishments and rewards. Without knowing how much information people have about these factors, it is difficult to establish if punishments are varying with outcomes in a way that reflects moral luck, as opposed to just inferring hidden ac-

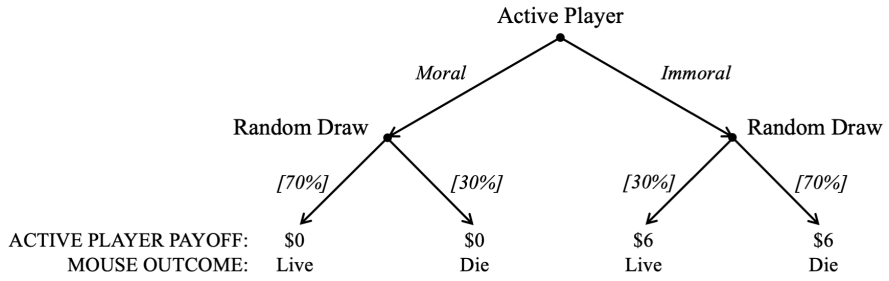
tion from outcomes.[1] One solution is to study decisions in controlled environments, where the researcher causes outcomes to vary in ways that are plausibly or explicitly due to chance (see Robbenholt, 2000; Martin and Cushman, 2016; Gurdal et al., 2013; Brownback and Kuhn, 2019), but using such artificial environments raises the difficulty of having real, consequential moral decisions. Having consequential decisions may be important for recruiting key mechanisms in a realistic way, e.g., eliciting strong emotions. Such realism is desirable for assessing how strong the phenomenon is, e.g., in the face of interventions designed to mute emotions and debias behavior. This study uses a framework that combines both clean identification of moral luck, with consequential moral choices that are a matter of life and death for a third party (a mouse).

2.2 Experiments

In a first stage of our experiments, shown in Figure 1, subjects in the role of active players make a choice between two lotteries, denoted the moral lottery and the immoral lottery, where outcomes are consequential in that they involve life or death for a third party. Specifically, the immoral lottery involves a 70% chance that a mouse dies, and a 30% chance that a mouse is instead rescued from death. The immoral lottery gives the active player $6 for themselves, regardless of the outcome for the mouse. The moral lottery, by contrast, involves only a 30% chance of death for the mouse, and a 70% chance that it is rescued, but gives the active player no money. An active player who chooses the immoral lottery thus indicates a willingness to increase the risk of death for the mouse, in order to achieve personal financial gain, whereas choice of the moral lottery reflects a willingness to sacrifice personal gain, in order to reduce likelihood of death for the mouse.

Our study uses the mouse paradigm developed in Falk and Szech (2013), where a key feature is that the population of mice used will be killed by default, in the absence of intervention through the study, and thus the scientific study can only improve welfare for the mice. The mice in question are ordinary laboratory mice, bred by a company for, e.g., medical research, but slated to be euthanized by the company to do lack of demand. If it is determined in our study that a mouse should be rescued, our research money is used to purchase one of these "surplus mice" from the company, and allow the mouse to live out the rest of its natural life in a hygienic environment with other mice.

1. If actions that influence the probabilities of good or bad outcomes are unobserved, or observed only with noise, then outcomes become informative signals of the intentions and actions of the actor, and thus the informativeness principle entails varying punishments and rewards with the outcomes to some extent. Without knowing exactly what decision makers believe, it is difficult to assess if punishments and rewards vary with outcomes to the optimal degree.



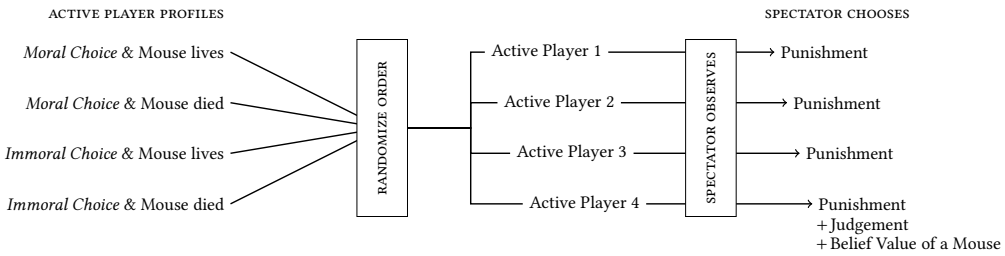
Notes: The active player first chooses one of two options, shown in the figure as *moral* or *immoral*, although more neutral, factual labels "option likely live" and "option likely die" were used with subjects. The moral choice leads to a subsequent random draw with a low probability of death for the mouse, 30%, and gives the active player no money regardless of what happens to the mouse. The immoral choice leads to a random draw with a high probability of death for the mouse, 70%, and gives the active player $6 regardless of what happens to the mouse. Note that the default for such surplus laboratory mice is to be killed, so the study is rescuing mice.

Figure 2.2.1. Stage 1 of experiment: Active player choices and outcomes

The first stage of our study also elicits traits and judgements of the active players. Specifically, we measure an active player's "value of the life of a mouse" using a question asking how much they would need to be paid, in order to allow a mouse to die for sure. In addition, the study measures active players' judgements about, e.g., the morality of their own choice, and whether they see themselves as a good person, after learning what happens to their mouse. The active players also have an additional, "pending payment" of $12; how much of this they receive depends on the choices of spectators in stage 2 of our experiment. We use university students as active players (N=562).

In the second stage of our experiment, which was pre-registered, we recruit a large sample of US adults to participate in online experiments in the role of spectators; our main treatment, Treatment Main, has N=2,200. We explain the concept of surplus mice to spectators, and elicit their (hypothetical) value of a life of a mouse. As was pre-registered, our analysis focuses only on spectators who have more than a minimal value for mice, to eliminate those who might dislike mice and thus not view active players as facing a moral dilemma. Spectators are given an endowment of $6, and can choose how much of this to spend, in order to reduce the pending payment of an active player. As shown in Figure 2, our design matches a given spectator with a sequence of four active players, so that they see each possible combination of choice and outcome for the mouse. The order of seeing the different active players with different possible choices and outcomes is randomized across spectators, to address any possible order effects. Spectators make a choice of how much money to deduct from each of the four active players, knowing that only one of the four choices will be randomly selected to potentially be implemented. In this sense, our

design is an example of the "strategy method," where subjects make choices without knowing for sure which case will be realized. Spectators knew that multiple spectators might be matched to a given active player, in which case it would be randomly determined which spectator's choice was used to determine the active player's payoff. This design allows a within-subject analysis. It can thus can speak to individual heterogeneity in a tendency to condition punishments on random outcomes, as well as the robustness of such a tendency to making the different possible choices and outcomes of actors salient to the spectator.



Notes: The spectators see a sequence of four different active players, with each possible combination of choice and outcome. The order is randomized across spectators. For each active player, the spectator has $6 to spend on punishment, with each dollar spent deducting two dollars from a pending $12 payoff of the active player. Spectators are asked for judgements and beliefs about the fourth active player that they see. Spectators know that one of the four active players will be randomly selected, and their punishment choice in that case will affect their payoff and potentially the payoff of the active player.

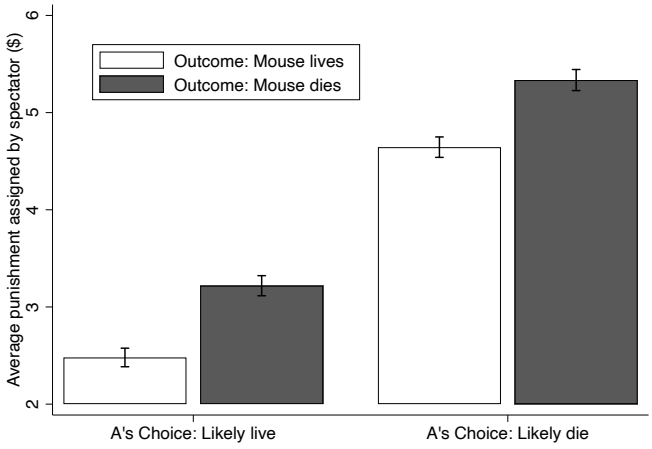
Figure 2.2.2. Stage 2 of experiment: Spectator punishment choices, judgements, and beliefs

The study also elicited judgements of the spectators about the fourth active player they saw, e.g., in terms of morality of the choice, and whether the active player was a good person. The elicitation asks only about the final active player a spectator saw, to reduce complexity of asking about all previous active players, and to focus on the one that was discussed most recently. We can compare across spectators how choices and outcomes affect judgements and emotions, because order is randomized. We also elicit incentivized beliefs of the spectator, about how the active player answered the question about value of the life of a mouse, paying the spectator for correctly guessing the money range indicated by the active player.

The rest of the study measures additional traits of the active player, and also assesses whether spectators exhibit moral luck in their judgements of hypothetical scenarios that span a range of contexts from crime, to politics, to economic interactions. Key traits that are measured include cognitive ability, captured by the cognitive reflection test (CRT) and a subset of Raven's progressive matrixes. We also ask about educational attainment. The questionnaire elicits agreement with the control principle, beliefs about the role of chance in determining outcomes like poverty in the US, and political affiliation and self-reported conservatism. Additional demographics include traits such as gender, age, and religion.

2.3 Results

Figure 2.3.1 shows our first set of results from Treatment Main, on whether there is moral luck in punishment choices. The figure shows average punishment levels by choice of the active player and outcome for the mouse, using all choices of spectators for a within-subject analysis. We see that punishments are on average significantly higher for active players who choose the immoral lottery, compared to those who choose the moral lottery, consistent with spectators sanctioning an immoral choice (OLS; s.e. clustering on spectator; $p < 0.001$). Punishments also vary significantly, however, with the outcome for the mouse, conditional on the active player's choice. For both the moral choice and the immoral choice, active players are punished significantly less if the mouse lives than if the mouse dies (OLS; s.e. clustering on spectator; $p < 0.001$, $p < 0.001$). Punishment choices thus violate the informativeness principle, in that active players are not being punished solely based on factors under their control. Results are similar and also statistically significant in a between-subject comparison, using only first choices of spectators (see Figure 2.A.1). This shows that the result is robust in the sense that it is not confined to within-subject contrasts. These findings raise the question whether moral luck in punishment reflects some alternative moral principle, or whether instead it is a mistake or bias.



Notes: Average punishment levels for each of the four cases. "A's choice" refers to Active Player's choice. Each spectator chooses punishment for all four cases so there are four observations per spectator (within-subject comparison). Figure shows standard error bars clustering on spectator.

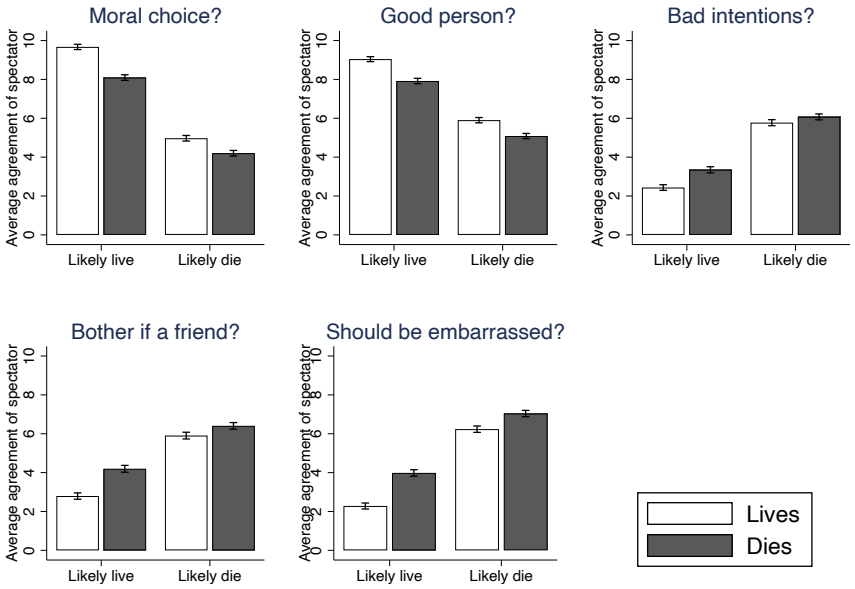
Figure 2.3.1. Punishment levels by active player choice and outcome in Treatment Main

Figures 2.3.2 and 2.3.3 explore one possible explanation (pre-registered), which is that punishments might vary with random outcomes because these influence judgements and beliefs about the nature of the active player, despite the fact that the

outcome conveys zero information above and beyond the observed choice. Figure 2.3.2 shows that for both the moral and immoral choice, the mouse dying causes spectators to judge the active player's choice as relatively more immoral, and to agree less that the active player is a good person. It also causes spectators to agree more strongly that the active player should be embarrassed, and that the active player had bad intentions. The impact of the mouse dying on each of these judgements is statistically significant, for both the moral and immoral choice, although interestingly, the effect of outcomes on most judgements is significantly stronger for the moral choice (see Table 2.B.1). Results are robust to controlling for spectator characteristics (see Table 2.B.2). We also elicited spectator beliefs about the active player's value of a mouse, because this might be a belief that is biased by whether the mouse dies, and potentially relevant for spectator punishment decisions. The belief measure also has the advantage that it can be incentivized for accuracy. Panel (a) of Figure 2.3.3 shows that the mouse dying significantly influences spectator beliefs about the active player's value of the life of a mouse (t-test; $p < 0.001$), particularly for the moral choice, where the effect is very large and individually significant (t-test; $p < 0.001$). The effect for the immoral choice is also positive, but smaller and not statistically significant individually (t-test; $p < 0.13$), and the effect is also significantly smaller than for the moral choice (OLS; $p < 0.001$). Finding a smaller effect for the immoral choice is consistent with the results on judgements. One explanation for these findings could be that spectators find the immoral choice as relatively more informative about the nature of the active player, and also recognize that it is consistent with only a relatively narrow range of values for life of a mouse, and this acts as a constraint on how much outcomes bias judgments and beliefs.

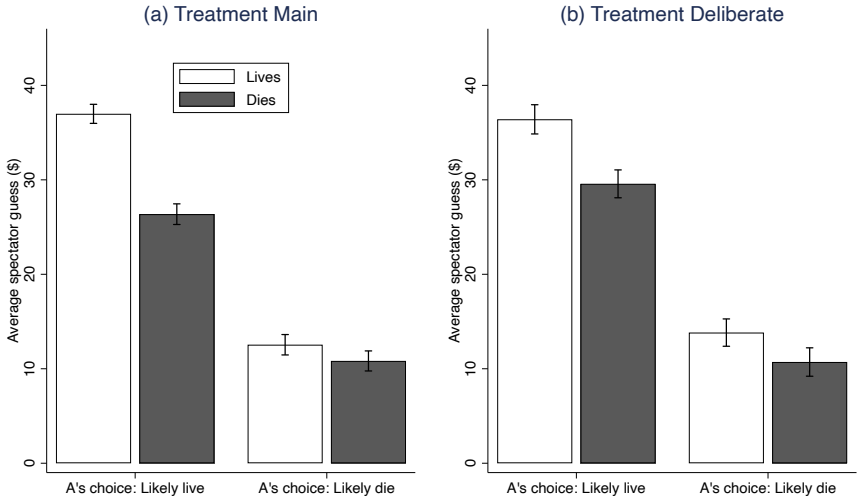
Despite conveying no information, the random outcomes influence judgements and even incentivized beliefs about the active player. If this is a mechanism underlying moral luck in punishment, one would expect punishment choices to be explained by variation in the judgements and beliefs. Table 2.3.1 presents evidence that this is the case. Columns (1) and (5) show that spectators who have less favorable judgements of the active player's choice or nature punish significantly more. Column (6) shows that incentivized beliefs about the active player's value of the life of a mouse are also significantly related to strength of punishment, with punishment decreasing in beliefs about how much the active player values a mouse. The results are robust to adding controls for other factors that matter for punishment choices, including choice and outcome of the active player, and spectator characteristics including own value of a mouse, income, education, gender, and educational attainment (see Table 2.B.3). These findings are consistent with the bias in perceptions of the active player, caused by the (uninformative) random outcomes, being a mechanism behind why punishment varies with random outcomes, but the evidence is correlational.

To provide evidence on whether the impact of random outcomes on perceptions of the active player is a causal mechanism explaining moral luck in punishment, we conducted a second treatment, Treatment Revealed Value (N=1,000). In this



Notes: Average agreement levels for each of the four cases. Each spectator judges one case so there is one observation per spectator (between-subject comparison). Figure shows standard error bars.

Figure 2.3.2. Judgements by choice and outcome in Treatment Main



Notes: Average incentivized guess about the active player's value of the life of a mouse. Each spectator makes a guess for one case so there is one observation per spectator (between-subject comparison). Panel (a) shows results from Treatment Main, Panel (b) from Treatment Deliberation. Figure shows standard error bars.

Figure 2.3.3. Spectator beliefs about the active player's value of the life of a mouse

Table 2.3.1. Punishment choices in Treatment Main as a function of judgements and beliefs

	Punishment ($)					
	(1)	(2)	(3)	(4)	(5)	(6)
Moral choice	-1.39***					
	(0.10)					
Good person		-1.26***				
		(0.11)				
Bad intentions			1.34***			
			(0.10)			
Bother if a friend				1.27***		
				(0.11)		
Embarassing					1.30***	
					(0.10)	
Belief about active player						-0.77***
						(0.10)
Constant	3.80***	3.84***	3.75***	3.77***	3.75***	4.02***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)
Observations	1441	1446	1443	1446	1446	1446
Adjusted R^2	0.127	0.100	0.112	0.096	0.105	0.037

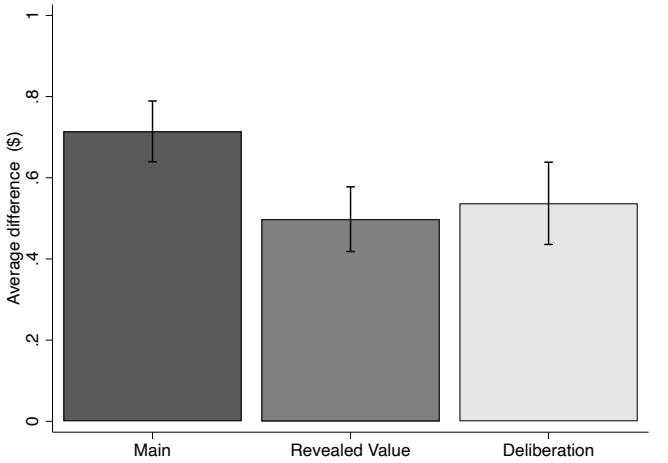
Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables include self-reported judgements about the active player given information about the fourth active player's choice and outcome for the mouse: Morality of active player's choice; active player is a good person; it would bother the spectator if active player was a friend; active player had bad intentions. Another independent variable is the spectator's incentivized guess about the active player's value of the life of a mouse, in dollars. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Each spectator makes judgements and reports beliefs for one case so there is one observation per spectator (between-subject comparison). Robust standard errors in parentheses.

treatment, spectators learned the active player's value of the life of a mouse, along with the choice and the outcome for the mouse. Each spectator was matched with four active players, two who chose the moral lottery and two who chose the immoral lottery. The active players who chose the moral lottery had different outcomes for the mouse, but had the same (high) value of a mouse, while the active players choosing the immoral lottery had different outcomes but the same (low) value of a mouse.[2] The key feature of the design is that value of a mouse was known to the spectator, and constant across active players with different outcomes, so there was no scope for random outcomes to influence beliefs. If part of the reason why punishment varies with outcomes in Treatment Main is the bias in beliefs about the active player value of a mouse, we would expect moral luck to be weaker in Treatment Revealed Value.

As shown in Figure 2.3.4, we find that punishment does, indeed, vary substantially less with the random outcome in Treatment Revealed Value compared to Treat-

2. The information conveyed about value of a mouse was calibrated to be line with priors conditional on choices. We used the modal guesses of spectators in Treatment Main, about values of active players choosing the moral or immoral lotteries, respectively, and selected active players with these values to use for the matching.

ment Main, a reduction of about 30 percent. This treatment difference is also statistically significant ($p < 0.05$; see Table 2.B.4). Moral luck is still, however, present and significant in Treatment Revealed Value, with punishment significantly stronger for cases when the mouse dies (t-test; $p < 0.001$). This is not unexpected, given that Treatment Revealed Value only eliminated the bias in a single aspect of how the active player is perceived, among a bundle of different types of judgments that are influenced by random outcomes and all seem to matter (correlationally) for punishment. Indeed, we find that judgements about the active player are still significantly skewed by random outcomes in Treatment Revealed Value, and these effects are not significantly different from in Treatment Main (see Table 2.B.5).



Notes: Average difference in punishment of die minus punishment of live. Each spectator chooses for all four cases so there are four observations per spectator (within-subject comparison). Figure shows standard error bars clustering on spectator.

Figure 2.3.4. Punishment of die minus punishment of live: Average difference by treatment

In another treatment, Treatment Deliberation, we investigate whether moral luck is robust to encouraging deliberative rather than intuitive thinking. We prime individuals to deliberate, through an essay asking about times when deliberation lead to good decisions, and intuition to bad, and also require a minimum time of 30 seconds to assign punishments, make judgements, and form beliefs. This treatment is based on previous approaches to encourage deliberative rather than intuitive decision making (Rand et al., 2012; Gino et al., 2008). If moral luck is weakened in this condition, it would suggest that violations are at least partly due to a mechanism of intuitive judgements, that are swayed by salient random outcomes when decisions are made quickly and spontaneously.

Figure 2.3.4 shows that encouraging deliberation does have a directional effect of reducing moral luck, leading to less variation in punishment with random outcomes, but the difference relative to Treatment Main is not statistically significant ($p < 0.16$; Table 2.B.4). Indeed, moral luck in punishment is still significant within Treatment Deliberation ($p < 0.001$; Table 2.B.4), and we also see signs of the bias mechanisms identified in Treatment Main, albeit somewhat weaker. Panel (b) of Figure 2.3.3 shows that beliefs about the active player's value of a mouse are significantly influenced by the mouse dying in Treatment Deliberation, overall (t-test; $p < 0.02$) and for both the moral and immoral choices individually (t-tests; $p < 0.01$, $p < 0.04$), and the effect is weaker for the immoral choice like in Treatment Main, although this difference is not statistically significant in Treatment Deliberation (OLS; $p < 0.22$). Overall, the impact of the mouse dying on beliefs is directionally weaker in Treatment Deliberation compared to Treatment Main, but the difference is not statistically significant (OLS; robust s.e. clustering on spectator; $p < 0.12$). We also see that outcomes significantly influence judgements in Treatment Deliberation. As for beliefs, the effects tend to be weaker than in Treatment Main, although the difference is significantly in the case of some judgements (see Table 2.B.5). Thus, it appears that deliberation may be directionally reducing moral luck by reducing, but not eliminating, the impact of outcomes on various judgements and beliefs. The fact that moral luck in punishment and the bias mechanisms are still present and significant in Treatment Deliberation suggest that the bias is relatively deeply-rooted and not easily corrected.

Taken together, our results show the existence of moral luck in punishment, and are consistent with the phenomenon being at least partly a bias, with random outcomes biasing judgements and beliefs about the active player, which in turn cause punishments vary with outcomes.

2.4 Additional Analysis

In additional analysis we explore some alternative explanations for why punishment might vary with outcomes, besides the bias that we identify, but find little support for these. One possible explanation is that individuals disagree with the control principle and have some other moral preference in mind that involves conditioning punishment on outcomes. We asked spectators directly whether they agree that people should only be punished for outcomes under their control, however, and find widespread agreement with this statement of the control principle, with the median individual expressing strong agreement, 9, on a scale from 0 (completely disagree) to 10 (completely agree). We also do not see a significant difference in the impact of the mouse dying on punishment, if we interact this with extent of agreement with the control principle (see Table 2.B.6; $p < 0.69$). Moral luck is also significant and strong even among those expressing complete agreement with the control principle.

These findings provide little support for an alternative moral principle as a driving force behind moral luck, and accord with one of the reasons moral luck as been seen as an ethical problem, that people will tend to agree with the control principle when asked, but violate this in practice (Smith, 1790; Nagel, 1979). Another explanation for why punishment varies with outcomes could be an imperfect understanding of the role of chance in our study, due to limited cognitive ability, or inattention to information provided, especially if this inattention is skewed towards noticing outcome information more than choice information. Working against such explanations, however, is the fact that spectators were required to correctly answer comprehension questions before making their choices. We also do not see a significant difference in the impact of the mouse dying on punishment, if we interact this with extent of cognitive ability, as measured by the cognitive reflection test, or Raven's progressive matrixes, or educational attainment (see Table 2.B.6; $p < 0.71$,$p < 0.89$,$p < 0.49$). We also find that spectators were attentive to the information provided. In an incentivized question at the very end of the study we asked spectators to recall the choice and outcome for the final active player they saw, and accuracy rates are quite high, about 85 percent, and essentially identical for choices and outcomes. Thus, relatively greater inattention to information about choices than outcomes does not appear to explain moral luck.

The bias we identify raises questions about what might be the deeper mechanisms that drive the bias; in exploratory analysis, we investigate three possible mechanisms – belief in a just world; hindsight bias or limited salience of counterfactuals; emotional impact of outcomes – and find some support for the final mechanism. The first two mechanisms would involve spectators viewing the bad outcome as more likely, if it occurs, and thereby potentially viewing the actor's choice as more immoral in that case. Belief in a just world is a type of motivated bias, such that people want to believe that bad things happen to bad people (Rubin and Peplau, 1975). Hindsight bias is a tendency for ex-post beliefs about the likelihood of an outcome to be greater than ex-ante, and has been hypothesized to reflect the fact that outcomes that occur are more salient than counterfactual outcomes (Roese and Vohn, 2012). A factor that works against these biases in our design, however, is the use of explicit probabilities. We also elicited spectator beliefs about the role of chance versus effort in determining inequality and poverty in the United States, as a proxy for belief in a just world, but find that the impact of the mouse dying on punishment is not significantly different depending on the extent of between belief in a just world (see Table 2.B.6;$p < 0.89$). The fact that we find strong moral luck in a within-subject design, where spectators make choices for all possible choices and outcomes, and counterfactuals are therefore salient, provides another indication that hindsight bias is not likely to be a key driver of the results. Lastly, we consider whether moral luck might be stronger for individuals who have stronger emotional reactions, suggesting a mechanism based on emotion. We first check whether outcomes affected emotions using a survey measure of self-reported emotions "about the fourth active player,"

on a scale from strongly negative to strongly positive, and find that the mouse dying significantly decreases positive emotions in the case of the moral choice (t-test; $p < 0.001$), and exacerbates negative emotions in the case of the immoral choice (t-test; $p < 0.001$). We hypothesize that spectators who experience stronger emotions in response to outcomes are those who care more about the outcome. As a candidate proxy for a trait of caring more about the outcomes, we take the spectator's own value of a mouse, and find that the impact of the mouse dying on negative emotions about the active player is, indeed, significantly stronger for spectators with a higher value of a mouse (OLS; $p < 0.002$). Turning to mechanisms for moral luck, there is a significantly stronger bias in beliefs about the active player's value of a mouse, for spectators who have a higher value themselves (OLS; $p < 0.001$), and a significantly stronger effect of the mouse dying on punishment for spectators who value a mouse more (Table 2.B.6; $p < 0.001$). One implication of these findings is that moral choices involving bad outcomes that are more emotionally upsetting may be more likely to generate moral luck. Another is that heterogeneity in moral luck may be partly explained by heterogeneity in how much punishers care about a given type of outcome.

Our within-subject design and use of a non-student sample allows us to investigate the prevalence of moral luck as a bias, as well as have meaningful variation in demographics and other correlates to explore whether the bias varies systematically across different segments of society. We find that exhibiting moral luck, defined as punishing more on average when the mouse dies than when the mouse lives, is the modal choice pattern in Treatment Main. Specifically, if we eliminate the 9 percent of spectators who do not exhibit moral luck because they never punish at all, we find roughly 43 percent exhibit moral luck, 36 percent zero moral luck, and 21 percent anti-moral luck. Thus, moral luck is prevalent but not universal. Anti-moral luck is less likely when individuals have higher cognitive ability, captured by the cognitive reflection test and Raven's progressive matrixes (OLS; $p < 0.001$, $p < 0.001$) and it is also smaller in magnitude than moral luck (t-test; $p < 0.007$), suggesting that this pattern reflects noise. We do not find significant differences in propensity to exhibit moral luck, or magnitude of moral luck, by gender, age, income, education, or political affiliation. Thus, the bias is found for individuals from across society. As noted above, one trait that does predict strength of moral luck is the spectator's own value of a mouse, pointing to caring about the outcome as a key moderator for moral luck in punishment.

Because we elicited judgments of active players about themselves, we can also explore an intriguing, additional question, which is whether moral luck is to some extent internalized by actors. Adam Smith and others have hypothesized that moral luck is internalized in this way, and one can also find examples from literature with this theme. For example, in ancient Greek tragedy, Oedipus kills his father in a roadside conflict, and marries his mother, without knowing their identities; when he later discovers what he has done, he blinds himself, and goes into exile, even though he

would presumably not have done had his vanquished opponent, and his wife, been unrelated to him. If random outcomes influence actors in how they judge themselves, and even potentially punish themselves (psychologically through feelings of guilt, or possibly through costly actions like "penance"), this would be a particularly striking form of moral luck, given that actors presumably have greater certainty about their own characters than external spectators.

We do find evidence of internalized moral luck for active players, although it differs in an interesting way from that of spectators. Specifically, active players judge their own immoral choice as significantly less immoral if the mouse lives than if the mouse dies (Table 2.B.7; $p < 0.03$). There is also suggestive evidence that actors who make the immoral choice change their view about being a good person based on the outcome, relative to a baseline assessment before their choice; the reduction in self-esteem if the mouse dies is marginally significant for individuals who have above median baseline self-image and therefore do not have a floor effect working against a reduction (Table 2.B.7; $p < 0.08$). Interestingly, however, we find an asymmetry, in that for active players there is little internalized moral luck for the moral choice. Active players view the moral choice as highly moral, regardless of the outcome, and also do not adjust their views of themselves as a good person (Table 2.B.7; $p < 0.9$, $p < 0.16$). These findings suggest that actors have a conviction that the moral action clearly indicates a good character, which cannot be shaken by having the mouse die, whereas they have more malleable views about the immoral action. This could potentially be motivated, if actors want to believe they are a good person; it may be possible to convince themselves of this in all cases, except for the immoral choice with the mouse dying. At the same time, we see that actors' feelings of embarrassment vary significantly with the outcome, for the moral as well as the immoral choice (Table 2.B.7; $p = 0.01$, $p < 0.01$). This suggests that actors anticipate that others may evaluate them based on outcomes for the moral choice, even if they themselves do not do so. This asymmetry in external versus internal moral luck that we find is in line with the type of tension hypothesized to arise in meritocracies, by Sandel (2019) and others, such that individuals who have had bad luck feel unfairly judged by others. Also, good or bad luck may have lasting influences on how individuals view themselves.

2.5 Discussion

Our findings have important implications for theories of human punishment behavior. Models of reciprocity theorize that individuals will engage in costly punishment of actions that cause harm. This can reflect a strategy of deterrence in repeated interactions, or it can arise as a heuristic or a preference for punishing those who would create harm by their actions, and manifest even in one-shot interactions (e.g., Trivers, 1971; Axelrod and Hamilton, 1981; Rabin, 1993; Levine, 1998; Dufwenberg

and Kirchsteiger, 2004; Falk and Fischbacher, 2006). A complicating factor in reality, however, is that both actions and chance often play a role in determining whether harm is caused. Our findings show that punishment behavior is influenced partly by actions, consistent with reciprocity theories, but also partly by random outcomes, something that cannot be explained by traditional models of reciprocity. Furthermore, we show that a key reason that random outcomes influence punishments is by biasing judgements and beliefs about the intentions of the actor. This implies lasting reputational effects of random outcomes, which could in turn lead to longer-run consequences for punishments that are also not explained by reciprocity models, e.g., in the form of future avoidance or ostracism of actors who are viewed as bad types due to previous bad luck with outcomes. Another novel prediction from our findings is that punishments may be particularly sensitive to outcomes when there is more limited information about intentions or characters of actors. Strangers, therefore, might be more subject to moral luck in how they are evaluated, compared to individuals' whose characters and reputations are well-known to evaluators (good or bad). Our findings call for modifying traditional models of reciprocity to allow for a bias in which evaluators wrongly infer about intentions and character from random outcomes.

The existence of moral luck in punishment also offers a new angle from which to theorize about how and why human punishment behaviors may have evolved. Evolutionary theories posit that human punishment behaviors may have played a crucial role in allowing the human species to sustain large-scale cooperation, by deterring actions that lead to harmful outcomes (e.g., Trivers, 1971; Axelrod and Hamilton, 1981; Henrich and Boyd, 2001; Fehr and Gaechter, 2002). Such theories, however, abstract away from the role of both actions and chance in determining harmful outcomes. One explanation for our finding that humans have a deep rooted tendency to condition punishment on outcomes, could be that it evolved to solve a problem of deterrence that arises in such cases, if actions are hard to observe (Gurdal et al., 2013). When actions influence the probability of harm, but are unobserved, outcomes are signals of actions, and optimal incentives involve conditioning punishment on the occurrence of harm. Our findings suggest that punishments are likely to be overly sensitive to outcomes, since they respond to outcomes even when actions are perfectly observable. But as a fast and frugal heuristic (Gigerenzer, 2004), moral luck could have been adaptive, if conditions with hard to observe actions were sufficiently frequent. One factor that may have also minimized the scope for moral luck to cause distortions in early societies, in cases when actions were observed, is the high frequency of repeated interactions. The result that having more information about actors reduces moral luck suggest that in early societies, with dense social networks and well-established reputations, distortions due to moral luck could potentially have been small, whereas in modern societies, where social networks are less dense, and there is less information about others' characters, moral luck has more scope to distort behavior.

The results in this paper complement previous empirical research on moral luck, and related concept of "outcome bias." The most common methodology has been hypothetical vignettes that vary whether an action is described as leading to more or less severe harm, and elicit moral judgements about the actor and views on appropriate punishment. Previous results are mixed, potentially due to issues of subjectivity in how subjects interpret scenario descriptions, especially interpretations of what different outcomes may signal about probabilities of harm, given that probabilities are typically implicit (for a survey and meta anlysis see, e.g., Robbenholt, 2000). Hypothetical measures also potentially encourage intuitive decision making and inattention, and likely attenuate emotional reactions, which may explain why asking subjects to decide rationally and deliberately has been found to significantly reduce moral luck in hypothetical vignettes (Gino et al., 2008), whereas in our setting with real outcomes and incentives we find persistent moral luck even with a relatively heavy-handed intervention. Previous research has found that incentivized beliefs about an actor can be influenced by random outcomes (Brownback and Kuhn, 2019), like we do, but we complement this finding by providing the first causal evidence that belief distortions due to random outcomes can cause moral luck in punishment behavior.

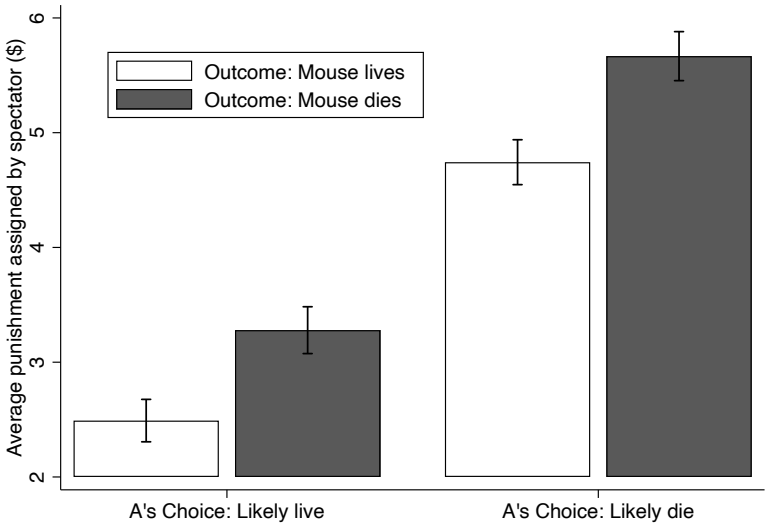
The findings of our study also add to the debate on whether moral luck is explained by a preference or moral principle, or is instead a bias. Theories of preferences over outcomes posit that individuals can care about outcomes per se, e.g., disliking inequality (Fehr and Schmidt, 1999). If the harm that results from an action leads to more inequality between the actor and another individual, punishment could be motivated by a desire to reduce inequality between these individuals. In our setting, however, it is unclear that inequality aversion applies, since when harm is caused the mouse is dead. Furthermore, we show that moral luck in punishment is driven by an impact of outcomes on perceptions of the actor, so the mechanism appears to be a form of reciprocity with biased beliefs, rather than inequality-averse preferences. It is also hard to explain our findings with adherence to a moral principle, since moral luck in punishment is driven by judgements and beliefs responding to outcomes that contain zero information. Instead, moral luck appears to be a bias. This does not mean, of course, that philosophical inquiry cannot make progress on seeking a moral principle that can justify conditioning punishment on random outcomes. Our findings do not answer the question of how people should make punishment decisions from a normative point of view, rather they shed light on the positive question of how people are making such decisions.

The fact that moral luck appears to reflect a bias, and has distortionary effects on deterrence, also suggests a potential value of interventions to reduce the bias. Our results suggest that effective interventions may include providing additional information about an actor, or encouraging deliberation, although this will probably not eliminate, the bias. One challenge for such interventions, however, is that evaluators may themselves be evaluated by the wider public, if the incidents they evaluate

are in the public view. To change behavior might therefore require an intervention to influence the public, not just the evaluator, which may be challenging. A recent example of such a situation could be the very public disqualification of the tennis player, Novak Djokovich from the 2020 U.S. Open. Djokovich hit a tennis ball in frustration towards the back of the court, and hurt a linesperson by hitting her in the throat. Video evidence shows that he was not looking where he was hitting, and if the path of the ball had been slightly different, no harm would have occurred. The rules of the tennis association call for disqualification for sufficiently severe recklessness, but leave it to officials to judge severity. In subsequent interviews, tennis officials agreed that Djokovich was not trying to hurt someone, and that if harm had not been caused, their decision would likely have been different. Since harm was caused, he was disqualified, and lost the $250,000 that he had earned for reaching the fourth round of the tournament. Tennis officials might have personally felt that the occurrence of harm was relevant for the decision, or they might have had doubts, but decided that the public would not be satisfied by anything less than disqualification.

In some cases, moral luck is seemingly codified in laws or rules within organizations, requiring evaluators to exhibit moral luck, raising the question whether there is a need for policy reform. An example is differences in sentencing guidelines, or rankings of severity of the crime, for attempted murder versus "successful" murder. Because this rule applies regardless of how hard the individual tried to commit murder, it seems that the key difference is whether, due to circumstances beyond the criminal's control, the murder attempt failed, and thus it exhibits moral luck. To the extent that legal judgements need to concur with notions of justice held by the general populace, and murder is an outcome with a particularly profound emotional impact, reforming legal codes to have the same punishment for attempted and successful murder may be difficult. Another argument against reform could be that it is too costly to determine the role of chance, and simpler to just adjust punishment based on whether outcomes occur, as these can be signals of good or bad intention. This seems contrary, however, to the notion of deliberation in legal judgements. Furthermore, in other areas of law, which involve civil rather than criminal offenses, the law is clear that severity of outcomes is not relevant for setting punishment. These seemingly contradictory ways of handling the role of chance in outcomes may reflect differences in severity of outcomes, and thus emotional reactions, and a tension between what seems rationally correct, and what feels correct.

Appendix 2.A Additional Figures



Notes: Average punishment levels for each of the four cases, with separate groups of subjects in each category. Figure shows standard error bars. Punishment is significantly higher when the mouse dies than when the mouse lives, for both the moral and immoral choice (OLS; $p < 0.004$, $p < 0.001$).

Figure 2.A.1. Punishment levels in Treatment Main using only first choices for between-subject comparison

Appendix 2.B Additional Tables

Table 2.B.1. Moral luck in judgments in Treatment Main

	Moral	Good	Embarrassing	Bother if friend	Bad intent.
	(1)	(2)	(3)	(4)	(5)
Immoral choice	-4.70***	-3.14***	3.34***	3.11***	3.95***
	(0.20)	(0.19)	(0.21)	(0.24)	(0.22)
Die	-1.58***	-1.13***	0.91***	1.40***	1.70***
	(0.18)	(0.17)	(0.20)	(0.24)	(0.22)
Immoral*Die	0.81***	0.31	-0.62**	-0.90***	-0.90***
	(0.29)	(0.27)	(0.31)	(0.34)	(0.33)
Constant	9.67***	9.04***	2.44***	2.79***	2.28***
	(0.10)	(0.11)	(0.12)	(0.16)	(0.13)
Observations	1441	1446	1443	1446	1446
Adjusted R^2	0.416	0.281	0.227	0.167	0.275

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variables are indicators for the immoral choice, outcome of mouse dying, and the interaction of these. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table 2.B.2. Moral luck in judgments in Treatment Main with controls

	Moral	Good	Embarrassing	Bother if friend	Bad intent.
	(1)	(2)	(3)	(4)	(5)
Immoral choice	-4.70***	-3.13***	3.33***	3.12***	3.92***
	(0.20)	(0.19)	(0.21)	(0.24)	(0.22)
Die	-1.62***	-1.16***	0.96***	1.44***	1.73***
	(0.19)	(0.18)	(0.20)	(0.24)	(0.22)
Immoral*Die	0.82***	0.32	-0.65**	-0.94***	-0.89***
	(0.29)	(0.27)	(0.31)	(0.34)	(0.33)
Constant	9.86***	9.31***	2.87***	2.92***	1.75***
	(0.46)	(0.44)	(0.50)	(0.58)	(0.54)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1441	1446	1443	1446	1446
Adjusted R^2	0.422	0.290	0.244	0.176	0.284

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variables are indicator for the immoral choice, outcome of mouse dying, and the interaction of these. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Controls consist of: The spectator's own value of a mouse; gender; age; income range; educational attainment. Robust standard errors in parentheses.

Table 2.B.3. Punishment choices in Treatment Main as a function of judgements and beliefs with controls

	Punishment ($)					
	(1)	(2)	(3)	(4)	(5)	(6)
Moral choice	-1.42***					
	(0.10)					
Good person		-1.28***				
		(0.10)				
Bad intentions			1.40***			
			(0.10)			
Bother if a friend				1.31***		
				(0.11)		
Embarassing					1.34***	
					(0.10)	
Belief about active player						-0.82***
						(0.11)
Constant	3.73***	3.86***	3.43***	3.61***	3.86***	3.38***
	(0.66)	(0.67)	(0.67)	(0.68)	(0.67)	(0.70)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1437	1442	1439	1442	1442	1442
Adjusted R^2	0.135	0.107	0.123	0.106	0.115	0.044

Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables include self-reported judgements about the active player given information about the fourth active player's choice and outcome for the mouse: Morality of active player's choice; active player is a good person; it would bother the spectator if active player was a friend; active player had bad intentions. Another independent variable is the spectator's incentivized guess about the active player's value of the life of a mouse, in dollars. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Controls consist of: Dummy variables for choice of the observed (fourth) active player and outcome for the mouse, with moral choice and lives as the omitted category; the spectator's own value of a mouse; gender; age; income range; educational attainment. Self-reported judgements and beliefs refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table 2.B.4. Treatment comparisons of moral luck in punishment

	Punishment ($)	
	(1)	(2)
Die	0.71***	0.71***
	(0.07)	(0.07)
T. Revealed Value	0.30**	0.29**
	(0.13)	(0.13)
T. Deliberation	-0.23*	-0.23*
	(0.14)	(0.13)
Die*T. Revealed Value	-0.22**	-0.22**
	(0.11)	(0.11)
Die*T. Deliberation	-0.18	-0.18
	(0.13)	(0.13)
Constant	3.56***	3.65***
	(0.08)	(0.34)
Controls	No	Yes
Observations	12120	12116
Adjusted R^2	0.007	0.010

Notes: OLS regressions. Dependent variable is punishment in dollars. Die is an indicator for the outcome of the mouse dying, measuring the differential punishment when the mouse dies versus when the mouse lives, averaged across the cases of moral and immoral choice. T. Revealed Value and T. Deliberation are treatment dummies, respectively. Controls include: The spectator's own value of a mouse; gender; age; income range; educational attainment. Each spectator chooses punishment for all four cases of active player choice and outcome so there are four observations per spectator (within-subject comparison). Robust standard errors in parentheses, clustering on spectator.

Table 2.B.5. Treatment comparisons of moral luck in judgements

	Moral	Good	Embarrassing	Bother if friend	Bad intent.
	(1)	(2)	(3)	(4)	(5)
Die	-1.38***	-1.12***	1.43***	1.09***	0.75***
	(0.18)	(0.16)	(0.19)	(0.18)	(0.17)
T. Revealed Value	-0.33	-0.28	0.16	0.28	0.21
	(0.22)	(0.18)	(0.22)	(0.22)	(0.20)
T. Deliberation	-0.21	-0.17	0.23	0.12	0.18
	(0.22)	(0.19)	(0.22)	(0.22)	(0.21)
Die*T. Revealed Value	0.49	0.44*	-0.23	0.04	-0.23
	(0.31)	(0.26)	(0.31)	(0.31)	(0.29)
Die*T. Deliberation	0.55*	0.52**	-0.59*	-0.18	-0.44
	(0.31)	(0.26)	(0.32)	(0.32)	(0.29)
Constant	7.47***	7.58***	4.13***	4.25***	4.00***
	(0.13)	(0.11)	(0.13)	(0.13)	(0.12)
Observations	3008	3030	3030	3030	3027
Adjusted R^2	0.025	0.021	0.028	0.021	0.007

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variable Die is an indicator for the outcome of mouse dying, T. Revealed Value and T. Deliberate are treatment dummies, respectively. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table 2.B.6. Moral luck in punishment choices in Treatment Main as a function of potential mechanisms

	Punishment ($)					
	punish	punish	punish	punish	punish	punish
Die	0.71***	0.72***	0.71***	0.72***	0.72***	0.51***
	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)
Agreement control principle	-0.11					
	(0.08)					
Die*Agreement control principle	0.04					
	(0.08)					
CRT score		-0.29***				
		(0.08)				
Die*CRT score		0.02				
		(0.07)				
Raven's IQ score			-0.33***			
			(0.09)			
Die*Raven's IQ score			-0.01			
			(0.07)			
Educational attainment				-0.01		
				(0.08)		
Die*Educational attainment				0.05		
				(0.07)		
Belief in a just world					-0.11**	
					(0.05)	
Die*Belief in a just world					-0.01	
					(0.05)	
Spectator's value of a mouse						-0.01
						(0.01)
Die*Spectator's value of a mouse						0.38***
						(0.09)
Constant	3.56***	3.54***	3.57***	3.56***	3.59***	3.71***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.21)
Observations	5784	5784	5784	5784	5772	5784
Adjusted R^2	0.008	0.011	0.014	0.007	0.009	0.009

Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables in include: Self-reported agreement with the control principle; CRT test score; Raven's IQ test score; spectator's own value of the life of a mouse. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Each spectator makes a choice for all four cases so there are four observations per spectator (within-subjects). Robust standard errors in parentheses, clustering on spectator.

Table 2.B.7. Moral luck for active players

	Immoral Choice			Moral Choice		
	Moral	Good	Embarrassed	Moral	Good	Embarrassed
	(1)	(2)	(3)	(4)	(5)	(6)
Die	-0.72**	-0.32*	0.94***	-0.05	0.15	1.78***
	(0.32)	(0.18)	(0.36)	(0.34)	(0.10)	(0.46)
Constant	4.23***	-0.43***	2.73***	8.38***	-0.12*	0.80***
	(0.28)	(0.14)	(0.29)	(0.21)	(0.06)	(0.21)
Observations	257	133	257	170	103	170
Adjusted R^2	0.02	0.01	0.02	-0.01	0.01	0.09

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) and (4) Choice was moral; (2) and (5) I am a good person (difference after-before choice); (3) and (6) I would be embarrassed if a friend learned my choice. Columns (2) and (5) consider only those subjects with above median self-esteem to avoid floor effect. Die indicates whether the mouse died. Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

References

Aristotle. 1984. *The complete works of Aristotle*. Vol. 2, Princeton University Press Princeton, NJ.

Axelrod, Robert, and William Donald Hamilton. 1981. "The evolution of cooperation." *science* 211 (4489): 1390–1396.

Bolton, Patrick, Mathias Dewatripont, et al. 2005. *Contract theory*. MIT press.

Brownback, Andy, and Michael A Kuhn. 2019. "Understanding outcome bias." *Games and Economic Behavior* 117: 342–360.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and economic behavior* 47 (2): 268–298.

Enoch, David, and Andrei Marmor. 2007. "The case against moral luck." *Law and Philosophy* 26 (4): 405–436.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and economic behavior* 54 (2): 293–315.

Falk, Armin, and Nora Szech. 2013. "Morals and markets." *science* 340 (6133): 707–711.

Fehr, Ernst, and Simon Gaechter. 2002. "Altruistic punishment in humans." *Nature* 415 (6868): 137–140.

Fehr, Ernst, and Klaus M Schmidt. 1999. "A theory of fairness, competition, and cooperation." *quarterly journal of economics* 114 (3): 817–868.

Gigerenzer, Gerd. 2004. "Fast and frugal heuristics: The tools of bounded rationality." *Blackwell handbook of judgment and decision making* 62: 88.

Gino, Francesca, Don A Moore, Max H Bazerman, et al. 2008. *No harm, no foul: The outcome bias in ethical judgments*. Harvard Business School.

Gurdal, Mehmet Y, Joshua B Miller, and Aldo Rustichini. 2013. "Why blame?" *Journal of Political Economy* 121 (6): 1205–1247.

Henrich, Joseph, and Robert Boyd. 2001. "Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas." *Journal of theoretical biology* 208 (1): 79–89.

Holmström, Bengt. 1979. "Moral hazard and observability." *Bell journal of economics*, 74–91.

Kant, Immanuel. 1784/2017. *Kant: The metaphysics of morals*. Cambridge University Press.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." *Review of economic dynamics* 1 (3): 593–622.

Martin, Justin W, and Fiery Cushman. 2016. "The adaptive logic of moral luck." *Blackwell companion to experimental philosophy*, 190–202.

Nelkin, Dana K. 2021. "Moral Luck." In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American economic review*, 1281–1302.

Rand, David G, Joshua D Greene, and Martin A Nowak. 2012. "Spontaneous giving and calculated greed." *Nature* 489 (7416): 427–430.

Robbennolt, Jennifer K. 2000. "Outcome severity and judgments of "responsibility": A meta-analytic review 1." *Journal of applied social psychology* 30 (12): 2575–2609.

Roese, Neal J, and Kathleen D Vohs. 2012. "Hindsight bias." *Perspectives on psychological science* 7 (5): 411–426.

Rubin, Zick, and Letitia Anne Peplau. 1975. "Who believes in a just world?" *Journal of social issues* 31 (3): 65–89.

Sandel, Michael J. 2020. *The tyranny of merit: What's become of the common good?* Penguin UK.

Smith, Adam. 1790/2010. *The theory of moral sentiments*. Penguin.

Thomas, Nagel. 1979. "Mortal questions." *Cambridge: CIP*,

Trivers, Robert L. 1971. "The evolution of reciprocal altruism." *Quarterly review of biology* 46 (1): 35–57.

Williams, Bernard. 1981. *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Chapter 3

Self-Serving Attributions in Belief Formation

Joint with Lasse Stötzer

3.1 Introduction

Perceptions about how just the society is differ greatly across individuals (e.g., Benabou and Tirole, 2006). Though being confronted with contradicting evidence in daily life, many individuals, in particular in the US, believe that "everyone can make it". Wealthy individuals think that their prosperity is predominantly rooted in their own abilities and not their privileges which may lead to anger, frustration and polarization within the society (Sandel, 2020).[1]

This paper aims to shed light on one mechanism that helps to explain individuals' persistently biased perceptions of how just a society is and of the own abilities by exploring the role of self-serving attributions in belief formation. More precisely, it studies whether a motive to have a positive self-image can yield lead (i) privileged individuals to attribute positive feedback disproportionally to their own abilities and less to the fact that they were advantaged, and (ii) disadvantaged individuals to attribute negative feedback disproportionally to the possibility that they were discriminated instead of to their own performance. In particular, the paper explores how far not only the belief about the own abilities but also a beliefs about an external fundamental, that is not directly linked to self-image, can be distorted to sustain a positive self-image.

The paper is conceptualized with the goal to shed light on basic belief formation processes. To be able to carefully study these in a controlled way, it employs a laboratory experiment. Exploring self-serving attributions necessitates a setting in which

1. Moreover, the belief how just the world is also has a direct impact on political preferences in societies (Alesina and Angeletos, 2005; Alesina and La Ferrara, 2005).

individuals care to uphold a positive self-view. We follow previous research (e.g., Eil and Rao, 2011; Zimmermann, 2020) and choose to give participants (noisy) feedback about their intelligence, an attribute known to be highly ego relevant. In the experiment subjects start by completing an IQ test on day one. On day two, subjects are informed that they are randomly placed in groups of ten and that all group members are ranked according to their performance in the IQ test. Subjects have to state their beliefs about their rank in the group. In a next step, they receive feedback about their performance in the IQ test. The feedback consists of three comparisons with randomly chosen group members. In contrast to related studies, the outcome of these comparisons not only depends on the performance in the IQ test but also on another factor, the *state of the world*. The state of the world is unknown to the subjects, randomly determined at the beginning of the experiment and can be either *just* or *unjust*. While in the just world feedback only depends on the IQ test performance, feedback in the unjust world also depends on the randomly assigned *type* of the subject. Subjects are either a *RED* or a *BLUE* type.[2] If the *state of the world* is unjust, RED types are privileged, meaning they will be ranked above a BLUE subject regardless of their true rank. Analogously, BLUE types are discriminated in the unjust world and will always be ranked below a RED group member. After the feedback in form of the three comparisons, we elicit subjects' posterior belief about being ranked in the upper half of the group (IQ test performance belief) and the likelihood of living in the unjust world (unjust world belief).

To present causal evidence on self-serving attributions, we run control conditions in which subjects observe the feedback of a randomly assigned person and, afterwards, state posterior beliefs about this person's IQ-test performance and likelihood of living in the unjust world. Thus, the key treatment variation is the elimination of the self-interested motives in our control conditions, as subjects should have no interest to paint an unknown and randomly assigned person in an overly positive light.

Our analysis considers those two out of the four feedback-type-combinations ("relevant cases") that yield clear predictions regarding our research questions: Red types that receive positive feedback and BLUE types that receive negative feedback.[3] To estimate causal effects, we compare the posterior beliefs relative to their Bayesian predictions (*relative IQ belief* and *relative unjust world belief*) in treatment and control conditional on test score.[4]

2. Subjects know their type and also the exact distribution of types in the group. There are always five BLUE and five RED types in a group of ten.

3. Following Eil and Rao (2011) positive feedback is defined as as winning all three comparisons and negative feedback as losing all three comparisons.

4. Conditioning on IQ test score allows a comparison between treatment and control, as case assignment is random in the control, but depends on the test score in the treatment condition. More precisely, in the treatment, case assignment depends on the type (RED/BLUE) and the feedback (pos-

If individuals attribute feedback in a self-serving way, RED types (privileged in the unjust world) should attribute positive feedback more towards their intelligence and less towards the possibility of living in the unjust world. Hence, with respect to our outcome measures, we should see higher relative IQ beliefs and lower relative unjust world beliefs in the treatment. Analogously, treated BLUE types (discriminated in the unjust world) should attribute negative feedback more to the possibility of being in the unjust world and less to their performance in the IQ test.

Unfortunately, our data collection process was interrupted by the COVID-19 pandemic. Due to this interruption, our current control sample consists of only 30% of the planned subjects. Therefore, all presented analyses that use a treatment-control comparison are preliminary and should be looked at with great caution.

We start our analysis by comparing how Bayesian the updating behavior is in the two relevant cases in the treatment. The advantage of this descriptive approach is that it does not depend on the completeness of the control group. We find that on average subjects in both cases behave relatively Bayesian. However, comparing the two cases more closely reveals an (insignificant) pattern, that is in line with what the mechanism of self-serving attributions would suggest. More precisely, relative to BLUE subjects with negative feedback, RED subjects who receive positive feedback seem to respect the strength of the signals more with respect to IQ and less with respect to living in an unjust world. However, these differences are insignificant.

In a next step, we compare treatment and control. Treated RED types who received positive feedback state a 0.134 (significant at 1 %) higher relative IQ test belief. However, the relative world belief does not differ between the treatment and control group (coefficient: -0.02). BLUE subjects with negative feedback state a significantly smaller relative IQ test belief (coefficient: -1.18) and a significantly lower relative unjust world belief (coefficent: -0.22) in the treatment group. Both estimates are completely contrary to our hypotheses. However, in particular the large effect on IQ is driven by extreme values of the control group, suggesting that the pattern may at least partly be driven by the small sample size. Taken all results together, our preliminary analysis does not suggest that individuals update in a self-serving way.

After the posteriors we elicited additional measures to illustrate potential consequences of self-serving attributions. However, all of these measures completely work through self-serving attribution towards the external fundamental (unjust world belief). As we do not find any sign for such a distortion, the additional measures cannot reflect the consequences of a biased world belief in the hypothesized way. Thus, we skip the discussion of these results in main part of the paper.

Our study relates to several strands of research in economics. First, there is a large strand of literature that looks at motivated beliefs. Within this literature, the

itive/negative). Type is randomly assigned. Feedback, however, is not completely random as it also depends on the performance in the IQ test.

studies that explore how the desire to uphold a positive self-view can explain the existence and persistence of overconfidence and the studies on short-term updating of beliefs about an ego-relevant characteristic in the presence of uncertainty are particularly related to this paper (Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005; Köszegi, 2006; Dana, Weber, and Kuang, 2007; Eil and Rao, 2011; Mobius, Niederle, Niehaus, and Rosenblat, 2011; Sharot, Korn, and Dolan, 2011; Burks, Carpenter, Goette, and Rustichini, 2013; Barron, 2016; Golman, Hagmann, and Loewenstein, 2017; Coutts, 2019; Exley and Kessler, 2019; Schwardmann and Van der Weele, 2019; Zimmermann, 2020).[5] We differ from these papers in two important ways. We add a second dimension of uncertainty that is traditionally absent from the literature, and we focus on self-serving attributions as a mechanism.[6]

Second, there is research that are concerned with the idea of self-serving attributions. Several studies highlight the consequences of self-serving attributions in the field of CEO and trading behavior, and financial markets (see, e.g. Daniel, Hirshleifer, and Subrahmanyam (1998), Gervais and Odean (2001), Hilary and Menzly (2006), Doukas and Petmezas (2007), Billett and Qian (2008), Li (2010), Libby and Rennekamp (2012), Kim (2013), and Hoffmann and Post (2014)). While these papers illustrate how attributions might shape individual decision making, they do not focus on belief formation. In contrast and most closely related to our study, the experimental investigation by Coutts, Gerhards, and Murad (2019) shows how individuals attribute noisy performance feedback that depends on a subject's and a teammate's performance. Our study uses the state of the world as an external fundamental as opposed to teammate's performance which offers the ability to understand attribution in a broader context without any strategic concerns and allows us to identify how an individual's attribution differs depending on the direction of influence of this external factor.

5. Another important context in which implications of motivated reasoning have been studied is moral behavior. Research shows that individuals distort their beliefs about how other people behave (Di Tella, Rafael and Perez-Truglia, Ricardo and Babino, Andres and Sigman, Mariano, 2015; Falk, Neuber, and Szech, 2020), charity performance (Gneezy, Keenan, and Gneezy, 2014; Exley, 2020), risk and ambiguity preferences (Haisley and Weber, 2010; Exley, 2016), preferences over fairness (Konow, 2000; Dana, Weber, and Kuang, 2007) or product quality (Chen and Gesche, 2017; Gneezy, Saccardo, Serra-Garcia, and Veldhuizen, 2020).

6. Research on self-serving attributions has a longstanding tradition in social psychology. Hastorf et al. (1970) famously noted that we "are prone to alter our perception of causality (...). We attribute success to our own dispositions and failure to external forces." Meta analyses of the psychology literature were conducted by Miller and Ross (1975), Zuckerman (1979), Arkin, Cooper, and Kolditz (1980), and Mezulis, Abramson, Hyde, and Hankin (2004). While Miller and Ross (1975) found evidence of biased attributions only in light of success, i.e. when positive feedback is disproportionately attributed to oneself, the more up to date and larger meta-analysis by Mezulis et al. (2004) found evidence on self-serving attributions for both success and failure. However, most of the studies are purely observational and do not allow for a causal identification of the attribution bias.

Third, there is research on the belief in a just world (e.g., Lerner, 1980; Alesina and Angeletos, 2005; Alesina and La Ferrara, 2005; Benabou and Tirole, 2006). We contribute to this literature by exploring a mechanism that may explain differences in perceptions of how just the world is.

The remainder of the paper is structured as follows. We first introduce the experimental design. Section 3 outlines the empirical strategy and derives our hypotheses. Section 4 and 5 present results. Section 6 concludes.

3.2 Experimental Design

3.2.1 Overview

A causal study of self-serving attributions in response to feedback requires an environment consisting of (a) a situation in which individuals are motivated to distort beliefs, (b) uncertainty about the underlying reason for the feedback such that individuals can make attributions, (c) (conditional) exogenous variation in the received feedback, and (d) a control condition in which the self-serving motives are erased. Our design accommodates all of these features.

Treatment. The experiment consists of two parts. In part one participants take an IQ test. In part two subjects' beliefs about the test performance are elicited. Subsequently, subjects receive noisy feedback about their performance on the test that can be caused by two factors. (i) On the one hand it can be the result from pairwise comparisons of the IQ test performance with three other randomly chosen participants. In this case, the feedback contains true information about the intelligence of the participants. (ii) On the other hand the feedback can stem from an 'external fundamental'. More precisely, each subject is assigned to a binary type, RED or BLUE, and lives in either a just or an unjust world. The (a priori equally likely) state of the world is unknown, but determines whether participants are evaluated equally (just world) or unequally (unjust world). If a subject lives in the just world, feedback only depends on true performance comparisons (see (i)). However, in the unjust world RED types are always ranked above BLUE types. Thus, if a BLUE type is compared to a RED type in the unjust world, the BLUE type looses the comparison. Importantly, subjects know whether they are potentially positively (RED) or negatively (BLUE) discriminated, i.e. which type they are, but they do not know the true state of the world. After the feedback, posterior beliefs about the IQ test performance and the state of the world are elicited as main outcome measures.

Figure 3.2.1 summarizes the resulting four different cases in the treatment. Subjects are randomly assigned to a type (RED or BLUE) and they receive either

		Feedback	
		Positive	Negative
Type	RED	**A**	C
	BLUE	D	**B**

Figure 3.2.1. Different Cases in the Treatment

positive or negative feedback.[7] To study self-serving attributions, we focus on the cases **A** and **B**. Only in these two cases subjects have the opportunity to boost (case A) or uphold (case B) a positive self-image by distorting their beliefs about the cause of the feedback.[8]

The use of an IQ test allows us to create an environment in which individuals are concerned about their self-image and therefore have a motive to distort their beliefs about their performance, satisfying the requirement (a). The introduction of a probabilistic state of the world that crucially determines feedback entails that subjects can attribute their feedback to either their IQ or the state of the world, or some combination of the two, thereby satisfying the requirement (b). Comparing each subject with three other randomly drawn participants yields exogenous variation of feedback conditional on the true performance in the IQ test. Thus, individuals with the same performance on the IQ test and the same type can receive both positive and negative feedback, fulfilling requirement (c). To explore whether people update differently in the absence of self-serving motives, a control condition is implemented (d).

Control. The key idea of the control condition is to mimic the updating task but without any motives to attribute the feedback in a self-serving way. Thus, the design is exactly the same like in the treatment condition with the only difference being that subjects receive feedback not about themselves but about another participant.

7. Following Eil and Rao (2011), we define feedback as positive if all three comparisons were won, i.e. the subject ranked first in all three comparisons, and negative if the subject lost, i.e. ranked second in , all three comparisons. In the Appendix we also alter the definition and look at subjects who received (mostly) positive or (mostly) negative feedback. Feedback is (mostly) positive if a subject won 2 or 3 of the comparisons. Analogously, feedback is (mostly) negative if the subject won 0 or 1 comparisons.

8. In cases B and C there is no uncertainty about the cause of the positive/negative feedback. For example, a RED type receiving negative feedback knows that it was due to her IQ test performance, independent of the world she lives in. In Section 3.3 there is a more detailed discussion about the cases relevant to our analysis.

3.2.2 Timeline

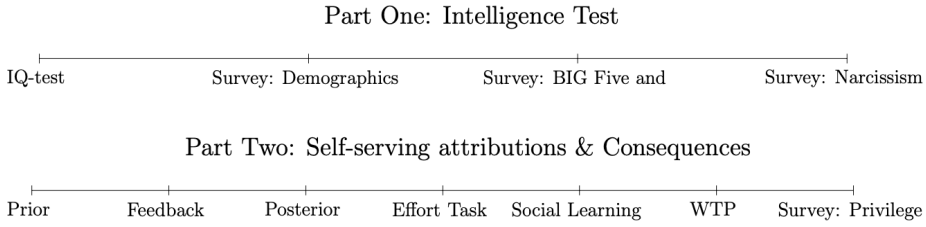


Figure 3.2.2. Timeline of the experiment

Figure 3.2.2 illustrates the timeline of the experiment. The experiment consists of two parts that span over two subsequent days. On the first day, participants did an IQ test and filled out surveys remotely. The main experiment took place on the second day and was carried out in the BonnEconLab. Subjects started by completing the 'self-serving attributions (SSA) segment' which contained three stages: (i) The elicitation of the prior beliefs, (ii) a feedback stage, and (iii) the elicitation of the posterior beliefs. Subsequently, we elicited further measures to get a better understanding of the mechanisms and consequences of self-serving attributions.

Part 1: Intelligence Test

IQ Test. On the first day, subjects had to complete carefully selected questions from a well-established intelligence test. In particular, they had to fill out three sections from the IST2000R IQ test measuring three distinct parts of intelligence: verbal, numerical, and spatial reasoning. On day onoee, subjects were not told that the questions were part of an intelligence test.

Surveys. Subsequently, subjects had to complete several questionnaires. Besides basic demographics questions, subjects had to fill out the 20-item IPIP-BFM-20 (BIG FIVE) questionnaire and the 16-item Narcissistic Personality Inventory.[9]

Part 2: Self-serving attributions (SSA) and Consequences

SSA: Prior. After revealing that the previous day's test measured intelligence, we informed subjects that they were randomly matched into a group with nine other people and that these nine other people had completed the identical intelligence test at an earlier time.[10] Based on the subject's and the other group members' per-

9. The demographics elicited include age, gender, field of study, highest degree, income, and risk seeking.

10. A week prior to the first session, we ran a small lab experiment to construct the values for the reference groups.

formance in the IQ test, we calculated a ranking of the group members.[11] In the next step, we elicited subjects' beliefs about their rank in the group: First, we asked subjects to state the likelihood that they are ranked in the upper half of the group. Second, to receive the full distribution, subjects had to estimate the likelihood of each of the ten positions in the ranking. Incentive compatibility was ensured using the quadratic scoring rule.[12]

SSA: Feedback. After the elicitation of the priors, the feedback stage of the experiment followed. Subjects were provided with noisy feedback about their performance in the IQ test. Following Eil and Rao (2011), a computer randomly selected one of the nine group members and informed subjects whether they ranked above or below the group member. We repeated this procedure three times, such that each subject observed the outcome of three comparisons. In contrast to Eil and Rao (2011), the outcome of the comparisons depended on the rank of the subject, but also on the combination of the subject's *type* and the *state of the world*:

- RED types were *positively* discriminated (privileged) in the unjust world.
- BLUE types were *negatively* discriminated in the unjust world.

Thus, if a RED subject was compared to a BLUE subject and the *state of the world* was unjust, the RED subject always won the comparison, i.e. ranked above the BLUE subject, irrespective of the true rank of the BLUE subject. Analogously, a BLUE subject always lost against a RED subject in the unjust world. In all other cases, i.e. in the case of two subjects of the same type in the unjust world and in all cases in the just world, the person with the higher IQ test score always won against the subject with the lower score. As a consequence, all feedback contained information about both an individual's rank as well as the state of the world.[13]

SSA: Posterior. After the feedback stage, we elicited the posterior beliefs about the intelligence and the state of the world. More precisely, subjects had to estimate the likelihood of being ranked in the upper half of the group (IQ test performance belief) and the probability of living in the unjust world (unjust world belief).

Consequences. The remaining stages relate to potential consequences of self-serving attributions. Distorted beliefs may lead to non-optimal decisions and biased perceptions. The respective measures serve as secondary outcome measures.

11. The subject with the highest score is ranked at position one, the subject with the second-highest score is ranked at the second position, and so on. In case of a tie, the ranks were randomly allocated.

12. For more details on the incentive scheme see Appendix 3.E.

13. The unjust world affected the different types in an analogous but diametrical way. This allows us to present evidence on the question of which circumstances facilitate self-serving attributions. RED types faced a world that potentially favored them while BLUE types were discriminated.

Real-Effort Task This stage explores whether distorted beliefs about the external fundamental (state of the world) can affect individual decision taking. The real effort task was a slider task similar to the one in Gill and Prowse (2012). To earn additional money, subjects had to win a comparison with a randomly drawn person who completed the same exercise at an earlier time. The outcome of the comparison depended on the number of sliders the subject pulled to 500 (the range was between 0 and 1000) and, as before, on the type of the subject and the state of the world. Analogously to the feedback stage, BLUE types always lost against RED types if the state of the world was unjust. If two subjects of the same type were compared or the state of the world was just, the number of correctly finished sliders determined who won. Thus, the unjust world belief affected the chances that the exerted effort paid off.
Social Learning This stage investigates whether self-serving attributions can lead to biased perceptions of others. Subjects observed the feedback received by a different participant that lived in the same state of the world but was part of a different group of ten. RED types always observed a BLUE type that lost all three comparisons and BLUE types always observed a BLUE type that won all comparisons. After observing the other participsant's feedback, subjects had to state their beliefs about the probability that the other person is ranked in the upper half of her respective comparison group and the belief about the (shared) state of the world.
Willingness to Pay The last stage elicited the monetary willingness to pay to learn the true state of the world via a price list.

After a short questionnaire, subjects learned how much they earned during the experiment.[14]

Control Condition

The control condition followed the same timeline. The key difference was that subjects, after stating their prior beliefs, were informed that the remainder of the experiment no longer concerned them, but that they would instead take on the role of a different, anonymous person ('Person Z') and receive feedback of that person. This person was randomly chosen and participated in one the treatment. Except for the type, subjects knew nothing about Person Z. After observing the feedback, subjects in the control conditions had to state posterior beliefs about Person Z's intelligence and the state of the world.

All parts of the experiment were incentivized. In Appendix 3.E we describe the incentive scheme of each task in more detail. To ensure that all subjects fully understood the instructions, control questions had to be completed before the feedback

14. In this last questionnaire, we elicited information about the subjects' socio-economic status and his or her sexual and religious preferences to explore potential heterogeneous effects.

stage. Likewise, an attention check was included by asking subjects to repeat their feedback directly after having seen it.

3.2.3 Logistics

A total of $N = 387$ subjects participated in the laboratory experiments: 292 in the *treatment* and, as of yet, 86 in the *Control*. The treatment sessions took place in October 2019 and the control sessions were implemented in March 2020. All sessions were conducted at the BonnEconLab of the University of Bonn. Most of the subjects were students from the University of Bonn. We used the hroot online recruitment system (Bock, Baetge, and Nicklisch, 2014) and computerized the experiment using o-tree experimental software (Chen, Schonger, and Wickens, 2016). Subjects spent an average of 27 minutes answering the online part and, on the subsequent day, about 45 minutes in the laboratory.

There was virtually no attrition between day one and day 2. Only 5 of 392 subjects that finished the first day of the experiment did not show up the following day.

Due to the COVID-19 pandemic, we were not able to complete the control sessions. As a consequence, some of our analyses are heavily underpowered and therefore must be interpreted with caution. We will address this problem at the relevant stages of our paper.

3.3 Hypotheses and Empirical Strategy

3.3.1 Self-Serving Attributions

The primary goal of this paper is to present causal evidence on self-serving attributions. To do so, we first discuss for which type-feedback combinations we should expect motivated attributions. Subsequently, we describe our empirical strategy and conclude by stating testable hypotheses.

Relevant Cases

The key idea of the experiment is that people generally want to attribute positive feedback to their intelligence and negative feedback to the state of the world. However, not all feedback-type combinations in our experiment generate a possibility to do so. As shown in Figure 1, there are four different combinations of feedback and type. In cases C and D, there is no uncertainty about the cause of the feedback as negative (positive) feedback for a RED (BLUE) type must stem from the IQ test performance - independent of the state of the world. However, in cases A and B participants have the chance to make self-serving attributions: RED (privileged) types who receive positive feedback may attribute it disproportionally to their intelligence and less to the possibility that they were positively discriminated, i.e.

that they live in an unjust world. In contrast, BLUE (discriminated) types who receive negative feedback may attribute it 'too much' to potential discrimination and 'too little' to their intelligence. Thus, we focus the analysis on the following two cases:

Case A: RED types who received positive feedback.

Case B: BLUE types who received negative feedback.

Empirical Strategy

Measures. To provide causal evidence on the mechanism of self-serving attributions, we compare the updating behavior in the treatment with the behavior in the control group. As subjects in the control group receive feedback not about themselves but another unknown participant, prior beliefs about the intelligence differ between control and treatment. In the treatment group, subjects form priors about their own intelligence based on their perceived performance in the test. In contrast, subjects in the control group have no information about the IQ test performance of the randomly allocated other person, such that they have to assign uniform probability to each rank. To circumvent this problem, we calculate the Bayesian predictions for both posterior beliefs using the stated IQ prior and the probability to live in the unjust world and construct following variables:[15]

Relative IQ test belief: rel_IQ
The relative IQ test belief *rel_IQ* of individual i is defined as the stated posterior of being ranked in the upper half of the group relative to its Bayesian prediction:

$$rel\_IQ_i = \frac{Posterior\_IQ_i}{Bayes\_IQ_i}$$

Relative unjust world belief: rel_world
Analogously, the relative unjust world belief *rel_world* of individual i is defined as the stated posterior of living in the unjust world relative to its Bayesian prediction:

$$rel\_world_i = \frac{Posterior\_World_i}{Bayes\_World_i}$$

15. See Appendix 3.D for more details on the derivations. The prior belief about living in the unjust world is, as known by the subjects, 0.5

Looking at the stated posteriors relative to the Bayesian posteriors erases the problem stemming from unequal priors, thereby rendering the updating behavior in the treatment and control group comparable.[16]

Econometric Strategy. To detect differences between the treatment and the control groups, we estimate following ATE specifications by OLS for each relevant case separately:

$$rel\_IQ_i = \alpha + \beta_{IQ} * treat_i + \gamma * iq\_score_i + \delta * controls_i + \epsilon_i \quad (3.3.1)$$

$$rel\_world_i = \alpha + \beta_{world} * treat_i + \gamma * iq\_score_i + \delta * controls_i + \epsilon_i \quad (3.3.2)$$

where $treat_i$ indicates whether subject i was in the treatment or the control group. Controlling for the IQ test score *iq_score* fulfills two purposes. First, it establishes comparability of subjects in the treatment and the control group as it makes feedback conditional exogenous.[17] Second, it analogously deals with differential selection into the two relevant cases within the treatment group based on IQ test performance. Hence, conditional on test score, there is no difference between subjects in treatment and control and no difference between subjects in cases A and B. In further specifications, more controls are added to get a more precise estimate of the treatment effect.[18] We consider the two cases separately, as the hypothesized treatment effects are different.

The main identifying assumption is that subjects in the treatment and control groups do not differ systematically conditional on test score. The biggest challenge to this assumption is stemming from the fact that there was no random assignment to treatment and control.[19] For funding reasons, treatment took place prior to control. Thus, even though participants were drawn from the same pool of potential subjects, there may be differential selection based on the point of time the experiment took place. Table 3.A.1 evaluates the size of the threat by comparing subjects' characteristics in control and treatment conditional on test score. There is no sign that subjects differ.

16. There exists a debate about whether starting from different priors effects updating mechanically (e.g., Coutts, 2019). To the best of our knowledge, there exists no conclusive evidence if the position of the prior (e.g., flat prior vs. all weight on first three positions) in itself distorts the updating behavior in any systematic way.

17. By construction, smarter (less smart) participants in the treatment group are more likely to receive positive (negative) feedback and thus, to be part of case A (B). However, in the control group there is no selection into cases based on intelligence. Controlling for test score deals with this endogeneity as it achieves conditional exogeneity of the feedback.

18. More precisely, we add subjects' age, gender, education level, the field of study, risk-seeking, and income. Further, we add the scores resulting from the Narcissism and BIG FIVE questionnaires

19. However, there was conditional random assignment to the different cases as discussed above.

Hypotheses Self-serving attributions

Case A (RED + positive feedback):

Hypothesis 1.1. Subjects who are potentially privileged (RED types) and receive positive feedback disproportionately attribute the feedback to their performance in the IQ test and underestimate the role of the external fundamental, i.e. the possibility that they live in the unjust world:

Equation (1): Relative to the Bayesian prediction, subjects in the treatment group state a higher posterior for being ranked in the upper half than the subjects in the control group ($\beta_{IQ} > 0$).

Equation (2): Relative to the Bayesian prediction, subjects in the treatment group state a lower posterior for being in the unjust world than the subjects in the control group ($\beta_{world} < 0$).

Case B (BLUE + negative feedback):

Hypothesis 1.2. Subjects who are potentially negatively discriminated (BLUE types) and receive negative feedback disproportionately attribute the feedback to the external fundamental, i.e. the possibility that they live in the unjust world, and underestimate the effect of their performance in the IQ test:

Equation (1): Relative to the Bayesian prediction, subjects in the treatment group state a higher posterior for being ranked in the upper half than the subjects in the control group ($\beta_{IQ} > 0$).

Equation (2): Relative to the Bayesian prediction, subjects in the treatment group state a higher posterior for being in the unjust world than the subjects in the control group ($\beta_{world} > 0$).

3.3.2 Consequences

Real Effort Task

The real effort task explores whether self-serving attributions can lead to non-optimal individual decision taking as a consequence to the biased belief about the external fundamental (state of the world). As described in section 3.2, we designed the real effort task in a way that the profitability of exerting effort depends on the likelihood of living in the unjust world. Given the state of the world is unjust, RED types always win against BLUE types, which makes exerting effort less attractive for both types. Thus, believing that the state of the world is unjust should, for both types, reduce the incentive to exert effort. To test this, we estimate the following

specification by OLS for the two relevant cases:

$$\text{Effort}_i \;=\; \alpha + \beta_{RE} * \textit{Posterior\_World}_i + \epsilon_i \tag{3.3.3}$$

Hypothesis 2. Effort is negatively correlated with the belief to live in the unjust world ($\beta_{RE} < 0$).

Social Learning

This stage investigates whether self-serving attributions can even lead to biased perceptions of others. The key idea is that subjects in the two relevant cases aim to uphold their unjust world belief as it enables them to maintain their positive self-image. When being confronted with information that challenges their 'desired state of the world', subjects prefer to make strong inferences about others than adapting their views of the world. To show this, we let RED types observe the feedback of a BLUE type that lost all three comparisons and let BLUE types observe the feedback of a BLUE type that won all three comparisons. To see if subjects indeed change their world belief too little but instead make too strong inferences about the other person, we estimate the following specification by OLS:

$$\textit{Social\_learning\_rel\_IQ}_i \;=\; \alpha + \beta_{SL} * \textit{treat}_i + \gamma * \textit{iq\_score}_i + \delta * \textit{controls}_i + \epsilon_i \tag{3.3.4}$$

where $\textit{Social\_learning\_rel\_IQ}_i$ is the belief about the IQ test performance of the person whose feedback the subjects observed. Apart from the outcome variable, equation (4) equals our main specification equations (1) and (2).

Hypothesis 3.1. (Case A: RED + positive feedback) RED types significantly underestimate the person's IQ test performance ($\beta_{SL} < 0$).

Hypothesis 3.2. (Case B: BLUE + negative feedback) BLUE types significantly overestimate the person's IQ test performance ($\beta_{SL} > 0$).

Willingness to pay

Existing research on motivated reasoning has shown that information avoidance is another tool that individuals use to sustain an overconfident self-view. To test if subjects in our experiment avoid information that potentially challenge their world view, we elicited their willingness to learn the true state of the world at the end of the experiment. Learning or not learning the state of the world at the end of the experiment has no strategic component to it. Therefore, if we see that subjects in the treatment have a lower willingness to pay, we can conclude that they try to

avoid receiving information that could threaten their self-view. We estimate following specification by OLS for the subjects in the two relevant cases:

$$\textit{Willingness\_to\_pay}_i = \alpha + \beta_{WTP} * \textit{treat}_i + \gamma * \textit{iq\_score}_i + \delta * \textit{controls}_i + \epsilon_i \tag{3.3.5}$$

Hypothesis 4. Subjects in the treatment group have a lower monetary willingness to pay to learn the true state of the world ($\beta_{WTP} < 0$).

3.4 Descriptive Analysis

In this section, we follow the empirical strategy by Eil and Rao (2011) and explore how far subjects deviate from Bayesian updating in the two relevant treatment conditions. More precisely, we estimate the following two specifications by OLS for both relevant cases in the treatment:

$$\textit{Posterior\_IQ}_i = \alpha + \beta_1 * \textit{Bayes\_IQ}_i + \gamma * \textit{controls}_i + \epsilon_i \tag{3.4.1}$$

$$\textit{Posterior\_World}_i = \alpha + \beta_2 * \textit{Bayes\_World}_i + \gamma * \textit{controls}_i + \epsilon_i \tag{3.4.2}$$

where $\textit{Posterior\_IQ}_i$ and $\textit{Posterior\_World}_i$ denote the stated IQ test performance belief and the stated unjust world belief after feedback. $\textit{Bayes\_IQ}_i$ and $\textit{Bayes\_world}_i$ are the respective Bayesian posteriors.[20] A β_1 (β_2) equal to 1 would indicate that subjects' updating behavior is similar to the Bayesian prediction of the respective outcome. A coefficient smaller than 1 would indicate a conservative reaction and a coefficient larger than 1 would indicate an overreaction.

Case A (RED + positive feedback). Table 3.A.2 reports the results for equations (6) and (7). The IQ coefficient (β_1) is 0.99 and the unjust world coefficient (β_2) is 0.85. Both coefficients do not significantly differ from one, i.e. we cannot reject the hypothesis that the subjects behave like Bayesian updaters.

Case B (BLUE + negative feedback). Table 3.A.3 reports the findings for case B. The IQ test coefficient of the Bayesian prediction is 0.72 and significantly different from one, indicating that subjects are relatively unresponsive to the feedback when it comes to updating over their IQ test performance. The unjust world coefficient is 1.22. The behavior does not deviate significantly from the Bayesian prediction.

It is notable how close stated posteriors are to the Bayesian predictions. In three out of four updating tasks, one cannot reject the hypothesis that participants update

20. As described in Section 3, we calculate the Bayesian predictions for both posterior beliefs using the stated IQ prior and the probability of the unjust world. For more details on the derivations, see Appendix 3.D.

in a Bayesian way. Only, when being discriminated and receiving negative feedback, there is "too little updating" with regard to the IQ.

The relation between the β_1s of cases A and B are in line with our expectations as participants in case A seem to react stronger to feedback than participants in case B.[21] Comparing β_2 in cases A and B also confirms the expected relation as β_2 is larger in case B, i.e. when participants receive negative feedback. To delve into the comparison further, Table 3.4.1 looks at both cases jointly and regresses the outcome variables *Posterior_IQ*$_i$ and *Posterior_World*$_i$ on the interaction of case assignment and Bayesian prediction conditional on iq-score.[22] The results are mixed. One the one hand, the pattern is as expected. In case B participants are less sensitive with respect to IQ and more sensitive with respect to the unjust world belief than in case A, as the respective coefficients of the interaction terms are -0.31 and 0.36. On the other hand, none of the coefficients are significantly different from zero.

Hence, there is some descriptive evidence suggesting that people attribute in a self-serving way. However, the facts that in three out of four cases we cannot reject Bayesian updating and that the differences between cases A and B are not significant, make the picture less clear. These results are descriptive in their nature and may partly be driven by cognitive constraints triggered by other factors like, for example, differential cognitive processes due to the rather complex setup of our experiment. Therefore, to gather clean causal evidence on our research question, we turn to the comparison of the updating behavior between subjects in the treatment and control groups in the next section.

3.5 Results

In this section, we provide perliminary causal evidence on whether people attribute feedback in a self-serving way by comparing updating behavior in treatment and control.

Remark: As highlighted before, the data collection process is not yet completed. As a consequence, the control groups are much smaller than the treatment groups. Concerning Case A (RED + positive feedback), there are 56 subjects in the treatment but only 17 subjects in the control group. Similarly, in Case B (BLUE + negative feedback) there are 56 subjects in the treatment but only 15 subjects in the control group. Therefore, the results presented here should be seen as first suggestive evidence in contrast to a thorough and comprehensive investigation of the research questions.

21. This is also in line with findings in the literature. See for example Eil and Rao (2011).
22. Note that case assignment is exogenous conditional on iq-score.

Table 3.4.1. Comparison of Cases A and B: Are Subjects Bayesian?

	IQ test		World unjust	
	(1)	(2)	(3)	(4)
Bayesian Prediction	0.859***	0.998***	1.127**	0.842
	(0.158)	(0.179)	(0.471)	(0.558)
Case B	12.46	23.08	-1.989	-34.67
	(13.79)	(15.30)	(41.87)	(50.61)
Bayesian Prediction × Case B	-0.187	-0.311	-0.0750	0.355
	(0.169)	(0.189)	(0.588)	(0.704)
Constant	-6.480	-48.54*	-7.725	22.90
	(14.78)	(28.71)	(36.96)	(55.64)
IQ Score	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
R^2	0.760	0.818	0.159	0.308
Observations	112	110	112	110

Notes: This table explores whether correlations with the Bayesian predictions differ across the two relevant cases. In columns (1) and (2) outcome variable and Bayesian predictions concern IQ test performance, in column (3) and (4) outcome variable and Bayesian predictions concern the unjust world belief. Feedback is defined as positive when a subject won all three comparisons. Case B is a dummy that equals one if a participant belongs to Case B instead of Case A. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Self-Serving Attributions

Case A (RED type + positive feedback). Figure 3.5.1 plots treatment coefficients for both outcome variables, the relative IQ test performance belief *rel_iq* and the relative unjust world belief *rel_world* (equations (1) and (2)). It shows that the relative IQ test belief *rel_iq* is significantly higher in the treatment group (coefficient: 0.134). In contrast, the relative unjust world belief does not significantly differ between treatment and control (coefficient: -0.02). Table 3.A.4 additionally provides estimates when controls are added (columns 3 and 6), and iq-score is dropped (columns 1 and 4). The results are very similar across specifications.

Our hypotheses of self-serving attributions state that while people take more credit for the positive feedback, they should simultaneously understate the role the world (external fundamental) played for the received feedback. This is only partly confirmed. While subjects in our treatment give themselves disproportionate credit for the positive feedback, they do not make self-serving inferences over the external fundamental.

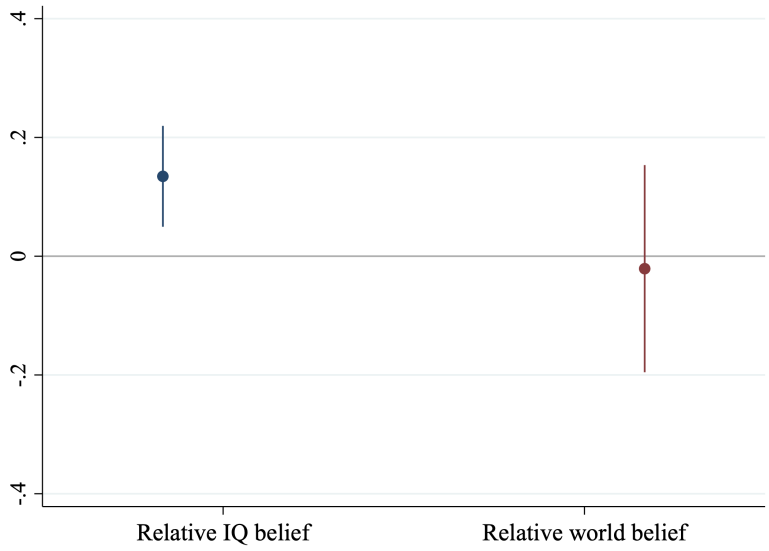


Figure 3.5.1. Treatment Effects in Case A

Notes: The figure plots the treatment effects for case A (RED type + positive feedback). The left coefficient denotes the treatment effect on the relative IQ test performance belief rel_{IQ}. The right coefficient denotes the effect on the relative unjust world belief rel_{world}. See Section 3 for more details. Error bars indicate 95% confidence intervals.

One possibility to get a larger comparison group and to check the robustness of our preliminary findings is to loosen the definition of positive feedback by altering its definition to having won at least two comparisons (instead of having won all three comparisons).[23] Table 3.A.6 runs the corresponding analyses. The results do not confirm the findings above. Neither the relative IQ-test performance belief nor the relative unjust world belief differ significantly between treatment and control. While this may be driven by the fact that the strength of the signal is 'too weak' under the alternative feedback definition as in this case also mixed feedback is regarded as positive feedback (having won 2 comparisons and having lost one comparison), or other factors, it does not support the result pattern above either. Thus, the preliminary findings should only be interpreted with great caution.

Case B (BLUE + negative feedback):. Figure 3.5.2 and Table 3.A.5 columns (2) and (5) show that BLUE types with negative feedback report significantly lower relative beliefs for both the IQ test performance *rel_iq* and the probability for being in the unjust world *rel_world*. For the relative IQ test performance belief the treatment

23. Note that even with this alternative definition the control group is still very small. Moreover, the imbalance between treatment and control stays the same as the alteration of thee definition yields more subjects in both cases.

coefficient is −1.18 (significant at 1 %), for the relative unjust world belief the coefficient is −0.22 (significant at 1 %). Table 3.A.5 shows that the results are similar when controls are added (columns 3 and 6), or iq-score is dropped (columns 1 and 4). When the definition of negative feedback is changed to having lost at least two comparisons (instead of having lost all three comparisons), we observe a similar yet less pronounced pattern. Table 3.A.7 shows that the treatment coefficients shrink to −0.99 (IQ) and −0.11 (world). The IQ estimate is still significant, but the world estimate is insignificant.

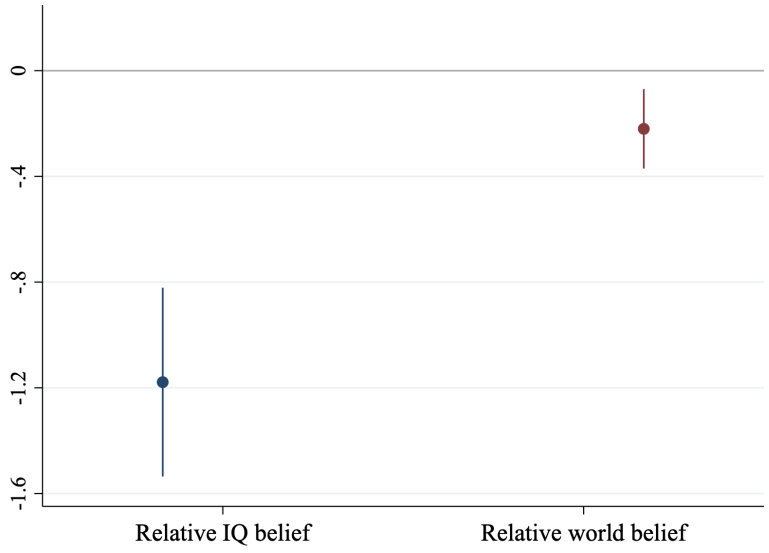


Figure 3.5.2. Treatment Effects in Case B

Notes: The figure plots the treatment effects for case B (BLUE + negative feedback). The left coefficient denotes the treatment effect on the relative IQ test performance belief rel_{IQ}. The right coefficient denotes the effect on the relative unjust world belief rel_{world}. See Section 3 for more details. Error bars indicate 95% confidence intervals.

The results are not in line with our hypotheses which predicted both coefficients to be significantly *larger* than zero. In particular, the estimate for IQ is very different from the predictions. One potential reason that might, at least partly, explain this result is founded in the control group. The average value of *rel_iq* is 2.29, i.e. subjects stated a posterior 2.3 times as large as a Bayesian would have done. To set the size in relation, the respective values in the other treatment and control groups are 1.12 (treatment, Case B), 0.93 (treatment, Case A) and 0.8 (control, Case A). Thus, subjects in the control group in case B were exceptionally non-Bayesian in their behavior. While one possibility certainly is that this captures the true average behavior in our control condition, the large values may also be driven by the very small size of the control group.

Altogether, the preliminary findings show no evidence for self-serving attributions, in particular to the external fundamental *state of the world*. Neither RED types that received positive feedback nor BLUE types that received negative feedback seem to distort their unjust world belief in a motivated manner. Whether these findings are founded in the small size of the control group, or whether people do not attribute feedback as hypothesized will be resolved when there are more subjects in the control groups. In Appendix 3.C, we discuss alternative mechanisms that may play a role instead of or in addition to self-serving attributions.

Consequences

The preliminary results do not show that subjects in the treatment group distort the unjust world belief in the hypothesized way. As all of our consequence measures build upon the existence of self-serving attributions towards the external fundamental, we do not include the analysis of the three consequence measures in the body of this paper. However, all analyses described in Section 3 can be found in Appendix 3.B.

3.6 Conclusion

This paper uses a laboratory experiment to study self-serving attributions in the case of two-dimensional uncertainty. After completing an IQ test, subjects received noisy feedback about their performance. The feedback depended on the actual performance on the IQ test, random noise and on whether the world was just or unjust. In the unjust world RED types were privileged and BLUE types discriminated. To learn from the feedback, subjects had to make attributions. We hypothesized that these attributions would be self-serving, meaning they would be made in such a way as to uphold or gain a positive self-image. More precisely, we expected that RED types that received positive feedback attributed the feedback too much to their own IQ test performance and too little to the possibility that the world was unjust. Likewise, we hypothesized that BLUE types attributed negative feedback too much to the possibility that they were discriminated and too little to their own intelligence. To provide causal evidence on self-serving attributions, we ran a control condition in which the motivational aspect was eliminated.

Due to the COVID-19 pandemic, we were not able to conduct all control sessions. As a consequence, our control group is relatively small and the preliminary results must be considered with caution. Our analysis explores two cases, privileged (RED) types that received positive feedback and discriminated types (BLUE) that received negative feedback. In neither case we find evidence that people make self-serving attributions. While RED types attribute positive feedback significantly 'too much' to their IQ test performance, there is no difference between treatment and control group with respect to the belief that the world was unjust. In the case of BLUE

types that receive negative feedback, the preliminary findings even suggest that they attribute the feedback in the opposite way as hypothesized, i.e. too much to IQ and too little the possibility of being discriminated. However, these results are mainly driven by the extreme values of the small control group.

The natural next step is to complete the control sessions. Having a sufficiently large control group will yield conclusive evidence on self-serving attributions. In case participants, contrary to our preliminary findings, attribute feedback in a self-serving way, it would be interesting to explore such behavior and its consequences outside the laboratory to assess its significance. Further possibilities for future research would be to explore the dependence of self-serving attributions on the salience of the potential causes, and to study the perseverance of the distortions.

Appendix 3.A Additional Tables

Table 3.A.1. Balance Check

	Treatment
Age	0.003
	(0.003)
Female	-0.021
	(0.046)
No Native	0.117
	(0.089)
Income	0.0
	(0.0)
Education	
Lower Secondary Education	-0.221
	(0.298)
Middle School	0.236
	(0.244)
Advanced technical certificate	-0.025
	(0.211)
High School	-0.01
	(0.115)
Other	-0.184
	(0.19)
F-Test	0.56
P-Value	0.74

Notes: The table reports the treatment coefficients of the balance checks. Dependent variables are age, share of females, native, income and education (categories). Each of these variables is regressed on the treatment dummy and iq-score. The respective dependent variable is listed in the left column. F-Tests of joint significance are calculated by regressing the treatment on all those variables and iq-score. Standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.2. Case A: Are Subjects Bayesian

	IQ test			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.911***	0.883***	0.994***	1.170**	0.937*	0.848
	(0.109)	(0.119)	(0.155)	(0.461)	(0.496)	(0.657)
Constant	1.687	-2.832	13.46	-15.44	25.86	8.712
	(9.548)	(11.92)	(26.76)	(30.54)	(44.96)	(70.99)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
P-Value (B.Pred.=1)	0.419	0.326	0.967	0.714	0.900	0.819
R^2	0.563	0.566	0.721	0.106	0.132	0.388
Observations	56	56	55	56	56	55

Notes: This table reports the respective correlations between the Bayesian prediction of RED types who received positive feedback and their stated IQ test performance (Column (1) and (2)) and unjust world belief (Column (3) and (4)). Feedback is defined as positive when a subject won all three comparisons. The 5th line (P-Value) tests whether the correlation coefficient of the Bayesian Prediction equals one. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.3. Case B: Are Subjects Bayesian?

	IQ test			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Pred.	0.712***	0.661***	0.718***	1.006***	0.907***	1.224**
	(0.0885)	(0.101)	(0.130)	(0.272)	(0.319)	(0.475)
Constant	14.05***	3.460	-27.97	-10.05	-10.78	-3.407
	(4.183)	(10.78)	(34.83)	(20.20)	(20.36)	(42.59)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
P-Value (B.Pred.=1)	0.00194	0.00138	0.0373	0.982	0.772	0.641
R2	0.545	0.555	0.757	0.202	0.207	0.463
Observations	56	56	55	56	56	55

Notes: This table reports the respective correlations between the Bayesian prediction of BLUE types who received positive feedback and their stated IQ test performance (Column (1) and (2)) and unjust world belief (Column (3) and (4)). Feedback is defined as positive when a subject won all three comparisons. The 5th line (P-Value) tests whether the correlation coefficient of the Bayesian Prediction equals one. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.4. Belief Updating in Case A (RED + Positive Feedback)

	Relative IQ test performance			Relative unjust World		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.138***	0.134***	0.143**	-0.00827	-0.0210	0.0298
	(0.0410)	(0.0426)	(0.0553)	(0.0844)	(0.0874)	(0.110)
Constant	0.795***	0.764***	0.603**	0.943***	0.812***	0.254
	(0.0359)	(0.112)	(0.271)	(0.0739)	(0.231)	(0.539)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.137	0.138	0.328	0.000135	0.00520	0.273
Observations	73	73	72	73	73	72

Notes: This table reports OLS estimates of subjects' relative posteriors on treatment for RED types that received positive feedback. Feedback is defined as positive when a subject won all three comparisons. The treatment dummy equals 1 if a subject received feedback about their own performance and 0 if a subject observed the feedback of a random other person. Columns (1) to (3) present results on the relative IQ test performance belief and columns (4) to (6) on the relative unjust world belief. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. Standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.5. Belief Updating in Case B (BLUE + Negative Feedback)

	Relative IQ-test performance			Relative unjust World		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-1.164***	-1.179***	-1.259***	-0.231***	-0.220***	-0.213**
	(0.173)	(0.179)	(0.184)	(0.0729)	(0.0753)	(0.0908)
Constant	2.288***	2.412***	2.522**	1.099***	1.002***	0.871*
	(0.153)	(0.397)	(1.023)	(0.0647)	(0.167)	(0.504)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.397	0.398	0.661	0.127	0.133	0.330
Observations	71	71	70	71	71	70

Notes: This table reports OLS estimates of subjects' relative posteriors on treatment for BLUE types that received negative feedback. Feedback is negative when a subject lost all three comparisons. The treatment dummy equals 1 if a subject received feedback about their own performance and 0 if a subject observed the feedback of a random other person. Columns (1) to (3) present results on the relative IQ test performance belief and columns (4) to (6) on the relative unjust world belief. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. Standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.6. Belief Updating in Case A (RED + Mostly Positive Feedback)

	IQ test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.00349	0.00277	-0.00191	0.0817	0.0772	0.0760
	(0.0452)	(0.0460)	(0.0531)	(0.0841)	(0.0857)	(0.0970)
Constant	0.960***	1.049***	0.970***	0.933***	0.870***	1.246**
	(0.0400)	(0.122)	(0.279)	(0.0744)	(0.227)	(0.510)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.0000489	0.00501	0.128	0.00767	0.00838	0.163
Observations	124	124	123	124	124	123

Notes: This table reports OLS estimates of subjects' relative posteriors on treatment for RED types that received positive feedback. Feedback is defined as positive when a subject won at least comparisons. The treatment dummy equals 1 if a subject received feedback about their own performance and 0 if a subject observed the feedback of a random other person. Columns (1) to (3) present results on the relative IQ test performance belief and columns (4) to (6) on the relative unjust world belief. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. Standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.A.7. Belief Updating in Case B (BLUE + Mostly Negative Feedback)

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.988***	-0.988***	-0.962***	-0.114	-0.112	-0.144*
	(0.139)	(0.140)	(0.154)	(0.0804)	(0.0796)	(0.0863)
Constant	2.046***	2.087***	1.617	1.046***	0.791***	1.222**
	(0.126)	(0.285)	(1.089)	(0.0727)	(0.162)	(0.612)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.328	0.328	0.501	0.0190	0.0478	0.316
Observations	105	105	103	105	105	103

Notes: This table reports OLS estimates of subjects' relative posteriors on treatment for BLUE types that received (mostly) negative feedback. Feedback is defined as negative when a subject lost all three or won only one comparisons. The treatment dummy equals 1 if a subject received feedback about their own performance and 0 if a subject observed the feedback of a random other person. Columns (1) to (3) present results on the relative IQ test performance belief and columns (4) to (6) on the relative unjust world belief. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. Standard errors in parentheses.* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 3.B Consequences

As stated in the body of the paper, we do not observe distorted unjust world beliefs in the treatment. Yet, all of our consequence measures are dependent on such attributions. Hence, we did not include our analyses on the consequences of distorted world beliefs in the main part of the paper. Nevertheless, we still present the results for the three measures.

Real Effort Task

After stating their posterior beliefs, subjects compete with a randomly chosen other subject in a real-effort task. The state of the world affected the outcome of the competition; When the world was unjust, RED types always won against BLUE types. Thus, for both the potentially privileged (RED) and potentially discriminated (BLUE) types the effort should decline with the stated likelihood for being in the unjust World. Fig-

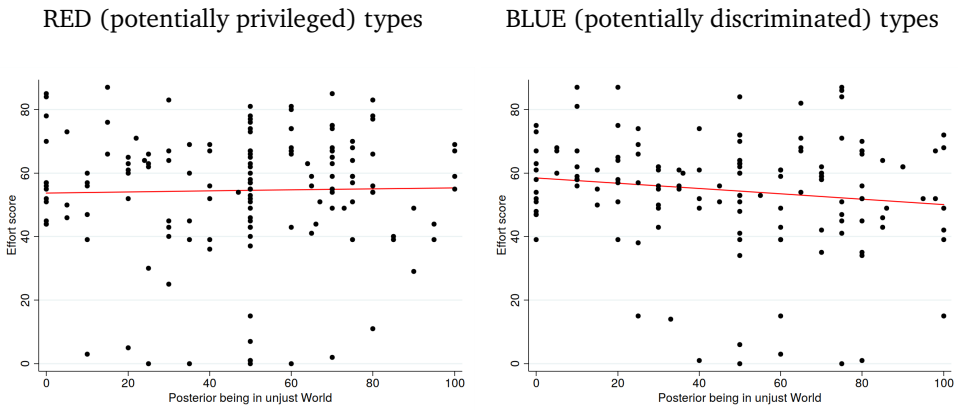


Figure 3.B.1. Effort Score and unjust World Belief

Notes: The two pictures plot the correlation between subjects' unjust world Posterior belief and the amount of sliders they finished in the real effort task (effort score). The left picture shows the relation for RED (potentially privileged) types and the right hand side picture the relation for BLUE (potentially discriminated) types.

ure 3.B.1 plots the correlation for the two types. We see that for RED subjects, the exercised effort does not change with the probability of being in the unjust world. In contrast there exists a weak negative correlation for BLUE subjects. Tables 3.F.1 and 3.F.2 report the corresponding results of the regression. Focusing on BLUE types, we see that a 10 point higher stated posterior leads to 0.8 less completed sliders (see Table 3.F.2 Column 1). This small and only weakly significant effect becomes insignificant as soon as we control for performance in the IQ-test. For RED types we do not detect any relation between unjust world belief and effort.

Learning about Others

We argued that subjects in both relevant cases uphold a positive self-view by respectively under- or overstating the likelihood for being in the unjust world. When observing the outcome of another person that threatens the own world view, e.g. a man that observes another privileged man that without any apparent skill holds a powerful position, individuals distort their assessment of the other person to maintain their own self-view. In our experiment BLUE types with negative feedback observe another BLUE type that won all three comparisons and lives in the same unknown world. RED types with positive feedback observe the feedback of a BLUE type that lost all three comparisons. Both signals should lead subjects to revise their distorted unjust world belief. But, as adapting the world belief would imply adapting the own self-view, we hypothesize that RED types would understate the performance of the other participant relative to the Bayesian prediction, whereas BLUE types would overstate it. Figure 3.B.2 illustrates the relative updating behavior in the treatment

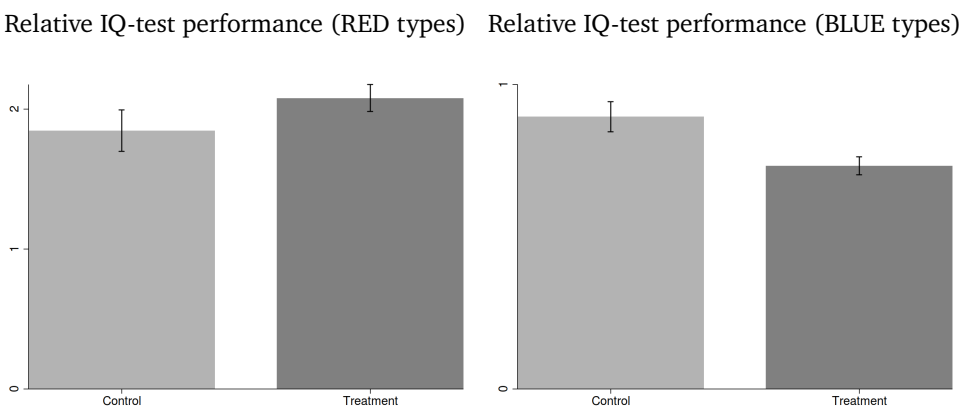


Figure 3.B.2. Learning about Others

Notes: After subjects observed feedback about a different participant, they had to assess the likelihood that this other participant is ranked in the upper half of her group. The left picture plots the average relative likelihood for the other participant to be ranked in the upper half (rel_{IQ_Other}) in the control and treatment group if the subject was of type RED (potentially privileged). The right hand side picture plots rel_{IQ_Other} in the control and treatment group if the subject was of type BLUE (potentially discriminated).The error bars indicate $+/-$ standard errors.

and control groups. Contrary to our hypothesis from Section 3.3 we observe that RED types assess the relative IQ-test performance of the other participant higher in the Treatment and BLUE types asses the performance lower in the treatment group. The corresponding numbers are reported in Table 3.F.3 and Table 3.F.4. For RED types the differences in the assessment of the other person between treatment and control groups are not significant. BLUE types in the treatment on average state a rel_{IQ_Other} of 0.733, i.e. the stated probabilities of the other participant being in the

upper half of her group are smaller than what a Bayesian person would state. In the control, subjects state a rel_{IQ_Other} of 0.895. The difference is significant at 5 %.

Willingness to Pay to Learn True State of the World

Next, we turn to whether subjects in our two cases avoid information about the true state of the world, even if avoiding information has no strategic advantage. To see this we compare the willingness to pay to learn the state of the world between treatment and control group. Table 3.F.5 reports our findings. We observe no differences in the willingness to learn the true state of the world. Subjects neither avoid information that could attack their motivated beliefs nor do they seem to be more curious.

Appendix 3.C Alternative Mechanisms

Due to the missing data in our control group, we have to look at the observed results from the previous section with great caution. In particular, the control group in Case B seems to completely contradict what we would expect. Nevertheless, our preliminary analysis revealed that subjects do not make self-serving attributions towards an external fundamental. In the following, we will discuss factors that can deepen our understanding of the updating behavior observed in the treatment. First, we will look into alternative mechanisms that could explain the subjects' behavior. The first alternative mechanism is *attribution to the noise or Good News vs. Bad News effect* and the second factor is *the role of the types in the unjust world*. Second, we study how *initial overconfidence* in the own ability affects the learning about an external fundamental.

Attribution to Noise or Good vs. Bad News Effect

A prominent hypothesis in the literature on short term updating with one dimension of uncertainty is that individuals react more strongly to favorable news than to unfavorable news (see, e.g. Eil and Rao (2011)). When individuals are confronted with negative feedback they surmise that the noise component of the feedback is to blame for the observed feedback, i.e. meaning they believe that the signal is just an unlucky random draw. Following the definitions from our results section, we compare the reaction of subjects that won all three comparisons (Good News) with subjects who lost all comparisons (Bad News). In the first step, we take all subjects in the treatment group together. To study how the behavior of the subjects depends on the sign of the feedback, we regress the stated posterior beliefs on the Bayesian predictions, a dummy for positive feedback, and the interaction between the two variables.[24][25] Figure 3.C.1 plots the reaction depending on whether subjects received good or bad news. Different slopes indicate that subjects reacted differently to pos-

24. The regression are as follows:

$$\textit{Post\_IQ}_i = \alpha + \beta_1 * \textit{Bayes\_IQ}_i + \beta_2 * \textit{good\_news} + \beta_3 * \textit{Bayes\_IQ}_i * \textit{good\_news} + \textit{Controls}_i + \epsilon_i$$
$$\textit{Post\_w}_i = \alpha + \beta_1 * \textit{Bayes\_w}_i + \beta_2 * \textit{good\_news} + \beta_3 * \textit{Bayes\_w}_i * \textit{good\_news} + \textit{Controls}_i + \epsilon_i$$

where *good_news* is a dummy equal to one if the subject won all three comparisons and zero if the subject lost all three comparisons.

25. Certainly one concern is that subjects who received positive feedback are inherently different from subjects who received negative feedback. Although the noise component in our feedback helps lessen this concern, we implement the following robustness checks to be able to present conclusive evidence: We run the same regressions comparing subjects that received mostly good (won 2 comparisons) and mostly bad news (won 1 comparison). See Tables 3.F.15, 3.F.16, and 3.F.17. We see that the results do not change. Further, when controlling for the feedback and the IQ-score, our results are robust, which indicates that the asymmetry does not depend on differences between subjects who received good and bad news.

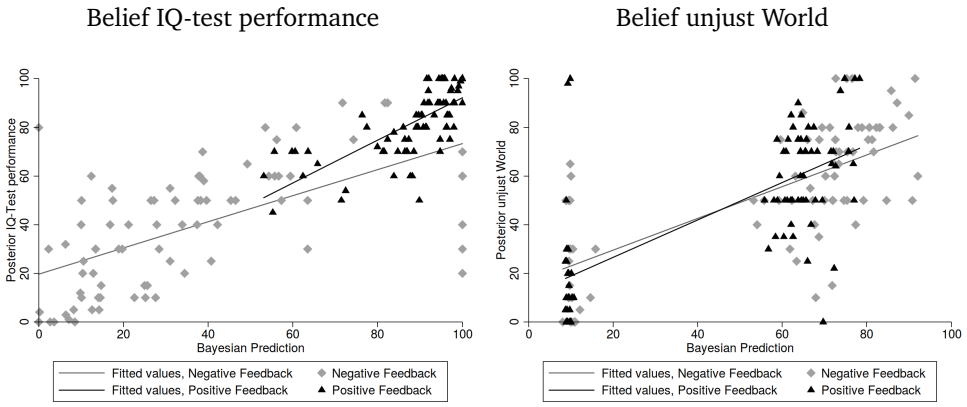


Figure 3.C.1. Good News vs. Bad News

Notes: The figures plot the subjects' posterior by the Bayesian prediction for the IQ-test performance Belief and the unjust World Belief. The data is split by the direction of the feedback.

itive and negative feedback. A steep slope indicates exaggerated responsiveness to the feedback. For the IQ-test performance belief, we see that the line for positive feedback is much steeper. This observation is confirmed by the significance of the coefficient of the interaction term in Table 3.F.9 column (1) to (4). In column (1) the interaction between the Bayesian prediction and the positive feedback dummy is 0.341 (significant at 5 %). Subjects perceive positive feedback to be stronger than negative feedback, i.e. when subjects receive positive feedback they seem to quickly adapt their self-evaluation while being rather unresponsive when facing negative feedback.

In the next step, we want to see if we observe this kind of reaction for both types. To do so, we split the sample and assess RED and BLUE types separately. We observe that the asymmetry in the reaction is purely driven by RED types. As shown in Table 3.F.10, the interaction term between IQ-test performance belief and Bayesian posterior for RED types is 0.821 (significant at 1 %). Meanwhile, the interaction for BLUE types is 0.0148 and insignificant (see Table 3.F.11). While potentially privileged (RED) subjects are only responsive to the received feedback when it is good news, the reaction of the potentially discriminated (BLUE) subjects does not depend on the direction of the feedback.[26] This raises the question if the randomly assigned type affects the updating of the subjects in general.

26. The effects do not change when adapting our definition of good and bad News to (mostly) positive and (mostly) negative feedback (see Table 3.F.15 for all types, Table 3.F.16 for RED types only, and Table 3.F.17 for BLUE types only).

Role of types in the unjust world

Our results on asymmetric updating revealed that while RED types react differently depending on the direction of the feedback, BLUE types do not. To fully understand the differences in updating behavior, we now focus on the role of the randomly given type. We address the following question: Do subjects react differently to the feedback depending on whether they are discriminated against or privileged in the unjust world?

To answer this question we regress the stated posterior beliefs on the Bayesian predictions, a type dummy, and the interaction between the two variables using all subjects from the treatment group.[27][28] Figure 3.C.2 and Table 3.F.12 report our findings. Figure 3.C.2 plots the reaction for the two types separately and a steeper slope indicates that subjects are more responsive to the feedback in their updating behavior. We observe that relative to the Bayesian predictions the updating behavior

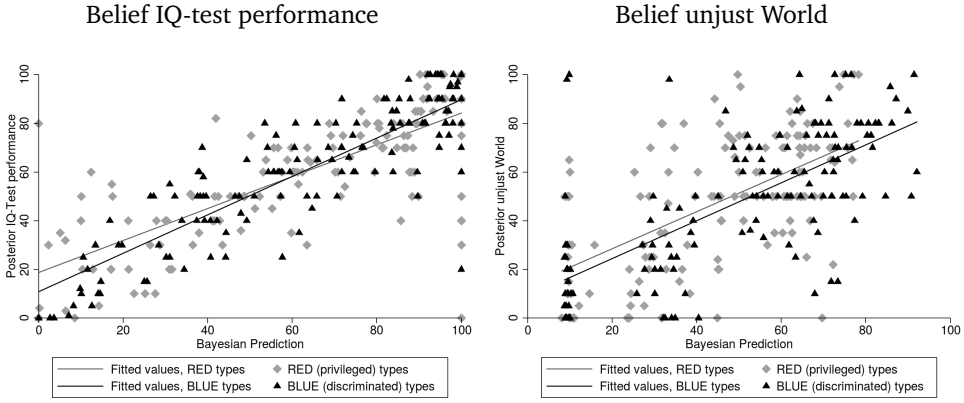


Figure 3.C.2. RED vs. BLUE types

Notes: The figures plots the subjects' posterior by the Bayesian prediction for the IQ-test performance Belief and the unjust World Belief. The data is split by the type of the subjects.

does not differ between types over the unjust world belief.[29] In contrast, updating behavior over the IQ-test performance belief differs significantly. As seen in the left picture in Figure 6, the slope of the BLUE types is steeper, indicating that BLUE participants are more responsive to the feedback. The interaction term between

27. This implies that we include all possible feedback x types combinations.
28. More specifically, we run the following regressions:

$$Posterior\_IQ_i = \alpha + \beta_1 * Bayes\_IQ_i + \beta_2 * type + \beta_3 * Bayes\_IQ_i * type + Controls_i + \epsilon_i$$

$$Posterior\_World_i = \alpha + \beta_1 * Bayes\_World_i + \beta_2 * type + \beta_3 * Bayes\_World_i * type + Controls_i + \epsilon_i$$

where *type* is a dummy equal to one if the subject is of BLUE type and zero if the subject is a RED type.

29. The slopes are identical but BLUE types seem to have a lower intercept.

type and Bayesian prediction in Table 3.F.12 column (1) to (4) adds numbers to this observation. In our specification without controls, we observe that the interaction term has a coefficient of 0.135 (significant at 5%). BLUE types have a lower intercept and a steeper slope, indicating that they generally respect the strength of the signal with regard to their IQ-test performance more than RED types.

We observe that subjects seem reluctant to adapt their IQ-test performance belief when receiving negative news. Looking at the two type-subsamples, we see that this effect is driven by RED (privileged) types. In line with this observation, we showed that BLUE types are generally more responsive to feedback. Again, this only holds true for the IQ-test performance belief. These effects stand out for two reasons: (i) they show that subjects' updating about the world is unfazed by the direction of the feedback and the type of the subject and (ii) randomly assigning subjects to a position of privilege (RED type) seems to affect their behavior. Privileged (RED type) subjects brush aside negative feedback and amplify the significance of positive feedback for their IQ-test performance. In contrast, BLUE types do not show this asymmetry and are generally more responsive to the feedback.

Initial Overconfidence

In the introduction, we raised the question of how individuals uphold overconfident self-views in the light of negative feedback. In the previous sections, we studied how self-serving attributions, asymmetric reactions to positive and negative feedback, and the role of the types in the unjust world can help explain this phenomenon. In this section, we investigate whether overconfidence affects reception to feedback both in terms of beliefs about intelligence and beliefs about the external fundamental.

A majority of subjects in our treatment sample state overconfident priors. The mean probability for being ranked in the upper half of the group before the subjects received feedback is 66.59 %. Further, the expected rank of the subjects is at least one rank lower (i.e. better) than the actual rank in approximately 75% of the cases. In the remainder of this section, we address the following two questions: Do overconfident subjects update differently from subjects with relatively correct beliefs? What is the impact of the initial bias on the unjust world Posterior? To answer these questions, we construct an overconfidence dummy variable (*over*).[30] As we seek to

30. The dummy variable equals one ($over = 1$) if a subject has an expected rank that is at least (or greater than) 1.5 ranks lower than the actual rank, i.e. a subject is ranked at position 4 and her expected rank is better than 2.5. The dummy variable is equal to zero ($over = 0$) if a subject's expected rank is less than 1.5 ranks away from their true rank. Thus, they have, relatively speaking, correct priors. One concern is that because of the margins in the group, subjects who performed relatively well can by definition never be overconfident (or at least it's highly unlikely). By controlling for IQ-score, feedback and our other measures, we aim to solve this problem.

study how overconfidence influences updating in an environment with self-serving motivations, we restrict our analysis to subjects in the treatment group.

Different Updating Behavior. In line with the analyses in the previous two subsections, we run several regressions with the stated posterior as our outcome variable and an interaction term between the Bayesian prediction and the *over* dummy as our variable of interest. A significant interaction term would suggest that overconfident subjects update differently, i.e. they are significantly less or more receptive to the feedback. Table 3.F.13 reports our findings. Focusing on the updating behavior over the IQ-test performance (Table 3.F.13 Column (1) to (3)), we observe weakly significant effects for the interaction between the Bayesian prediction and the overconfident dummy. The coefficient in the specification without controls is -0.168. This suggests that overconfident subjects react less strongly to feedback. Although insignificant, we observe the same pattern for the unjust world belief (3.F.13 Column (4) to (6)). Overconfident subjects seem to be more reluctant to adapt their (biased) initial belief.

Initial Bias. Recent theory papers by Heidhues, Kőszegi, and Strack (2018) and Hestermann and Yaouanq (2020) discuss how overconfident priors lead to distorted learning about an external fundamental. While our paper focused on biased updating to explain self-serving attributions, these two cited papers argue that the source of self-serving learning over an external fundamental is biased initial beliefs. In the context of our experiment this would imply that overconfident RED types, even when following Bayes Rule, end up with a lower unjust world belief than RED types that have accurate priors. Overconfident Bayesian BLUE types would end up with a significantly higher unjust world belief than a Bayesian BLUE subject with an accurate prior.
To present a first indicator of how overconfident priors distort the learning over an external fundamental, we regress the *over* dummy on the stated unjust world belief while controlling for the received feedback and the rank of the subject. We run this regression separately for the two types and only use subjects from the treatment group. We argue that given the identical rank and feedback, an overconfident RED type should end up with a lower unjust world belief. Analogously, a BLUE type should have a higher unjust world belief after receiving feedback. Table 3.F.14 displays the results. As hypothesized, overconfident RED types on the average state a -11.12 lower posterior. However, the observed effects are are only slightly significant, if at all. For BLUE types it is impossible to identify a clear tendency.

Taken together, we observe that overconfidence, in some cases, affects subjects' posterior beliefs. When it comes to the IQ-test performance belief, subjects who stated an expected rank that was 1.5 lower (i.e. better) than their actual rank are comparably unresponsive to the received feedback, which is to say they are reluctant to adapt their IQ-test performance beliefs. We further saw that for RED types who

had an identical rank and received identical feedback overconfidence leads to lower unjust world beliefs.

Appendix 3.D Bayesian Predictions

Using Bayes' Rule we derive our prediction for the IQ-test performance Belief. Let *upper_half* be the event that the subject is ranked in the upper half of her comparison group of ten. Let F denote the feedback the subject receives.

$$Bayes\_IQ = P(\text{upper half}|F) = \frac{P(F|\text{upper half}) * P(\text{upper half})}{P(F)}$$

where:

$$P(F|\text{upper half}) = P(unjust) * P(F|\text{upper half}, unjust) + P(just) * P(F|\text{upper half}, just)$$

$$P(F|\text{upper half}) = 0.5 * P(F|\text{upper half}, unjust) + 0.5 * P(F|\text{upper half}, just)$$

$$\begin{aligned} P(F) &= P(unjust) * P(F|unjust) + P(just) * P(F|just) \\ &= P(unjust)P(F|\text{upper half}, unjust) + P(just) * P(F|\text{upper half}, just) \\ &\quad + P(unjust) * P(F|\text{lower half}, unjust) + P(just) * P(F|\text{lower half}, just) \\ &= 0.5 * P(F|\text{upper half}, unjust) + 0.5 * P(F|\text{upper half}, just) \\ &\quad + 0.5 * P(F|\text{lower half}, unjust) + 0.5 * P(F|\text{lower half}, just) \end{aligned}$$

The Bayesian Prediction for the unjust World Belief:

$$Bayes\_World = P(unjust|F) = \frac{P(F|unjust) * P(unjust)}{P(F)}$$

where:

$$P(F|unjust) = P(unjust) * [P(F|\text{upper half}, unjust) + P(F|\text{lower half}, unjust)]$$

$$P(F|unjust) = 0.5 * [P(F|\text{upper half}, unjust) + P(F|\text{lower half}, unjust)]$$

$$\begin{aligned} P(F) &= P(unjust) * P(F|unjust) + P(just) * P(F|just) \\ &= P(unjust)P(F|\text{upper half}, unjust) + P(just) * P(F|\text{upper half}, just) \\ &\quad + P(unjust) * P(F|\text{lower half}, unjust) + P(just) * P(F|\text{lower half}, just) \\ &= 0.5 * P(F|\text{upper half}, unjust) + 0.5 * P(F|\text{upper half}, just) \\ &\quad + 0.5 * P(F|\text{lower half}, unjust) + 0.5 * P(F|\text{lower half}, just) \end{aligned}$$

Appendix 3.E Design - Incentive Scheme

The experiments spanned over two consecutive days and consisted of 6 payoff relevant components..The first incentivized component was subjects' performance in the IQ-test. Subjects received 10 cents for every right answer and in total they could earn up to 6 EUR. Next came subjects' prior beliefs about their performance in the IQ-test. Subjects stated the probability of being ranked in the upper half of their group of ten and the likelihood for each rank. Subjects could earn up to 2 EUR. One of the eleven beliefs was randomly chosen for payout. The third incentivized component was either the posterior beliefs about IQ test performance or about the state of the world. One of the two beliefs was randomly chosen for payout. In all belief elicitations, incentive compatibility was ensured by the quadratic scoring rule.[31] The next component is the real effort task, in which subjects could earn 4 EUR. The fifth component is inferences about a different person's performance on the IQ test. In particular, subjects observed the feedback of a different person and, based on this information, had to make inferences about the different person's performance in the IQ-test and about the external fundamental. Incentive compatibility was again ensured using the quadratic scoring rule. Last, using a price list we elicited subjects' willingness to pay to learn the true *state of the world*. At the end, one of the 21 price list choices was implemented.

31. The formula for the quadratic scoring rule for all beliefs (Priors and Posterior about IQ and external fundamental) was

$$Earnings = 2 - 2(I(true) - \frac{belief}{100})^2$$

where $I(true)$ is an indicator function. In the case of a subject's belief that she is in the upper half of the ranking, the indicator function takes value 1 if a subject is indeed in the upper half of the ranking and 0 otherwise.

Appendix 3.F Additional Tables Appendix

Table 3.F.1. Correlation unjust world belief and effort score - RED types

	(1)	(2)	(3)
Posterior unjust World	0.0168	-0.0375	-0.0241
	(0.0603)	(0.0607)	(0.0640)
Constant	53.73***	26.55***	61.50**
	(3.295)	(8.840)	(26.80)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.000524	0.0693	0.301
Observations	150	150	148

Notes: This table reports the correlation between the subjects' unjust World belief and the amount of sliders the subject pulled to 500 (effort score) for RED types. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.2. Correlation unjust world belief and effort score - BLUE types

	(1)	(2)	(3)
Posterior unjustworld	-0.0845*	-0.0426	-0.0416
	(0.0498)	(0.0489)	(0.0581)
Constant	58.53***	30.30***	31.36
	(2.775)	(8.014)	(25.83)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.0202	0.110	0.248
Observations	142	142	139

Notes: This table reports the correlation between the subjects' unjust World belief and the amount of sliders the subject pulled to 500 (effort score) for BLUE types. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.3. Learning about Others - RED type with positive feedback

	(1)	(2)	(3)
Treatment	0.232	0.186	0.158
	(0.193)	(0.199)	(0.248)
Constant	1.848***	1.374**	1.405
	(0.169)	(0.526)	(1.217)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.0199	0.0324	0.287
Observations	73	73	72

Notes: This table reports OLS estimates of subjects' relative learning about another participant's performance on treatment for RED types that received positive feedback. Feedback is said to be positive when a subject won all three comparisons. Treatment Dummy equals 1 if a subjects belongs to the treatment and 0 if a subject belongs to the control group. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.4. Learning about Others - BLUE type with negative feedback

	(1)	(2)	(3)
Treatment	-0.162**	-0.125*	-0.116
	(0.0637)	(0.0633)	(0.0789)
Constant	0.895***	0.582***	0.676
	(0.0565)	(0.140)	(0.438)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.0856	0.158	0.307
Observations	71	71	70

Notes: This table reports OLS estimates of subjects' relative Learning about another participant's performance on treatment for BLUE types that received negative feedback. Feedback is said to be negative when a subject lost all three comparisons. Treatment Dummy equals 1 if a subject belongs to the treatment and 0 if a subject belongs to the control group. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.5. Willingness to pay in treatment and control for the two cases

	(1)	(2)	(3)
Treatment	0.0217	0.0217	0.00614
	(0.0482)	(0.0484)	(0.0544)
Constant	0.0793*	0.0730	0.254
	(0.0425)	(0.104)	(0.342)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.00158	0.00161	0.173
Observations	130	130	129

Notes: This table reports OLS estimates of subjects' willingness to pay to learn the true state of the world on Treatment for the two relevant cases. Treatment Dummy equals 1 if a belongs to the treatment and 0 if a subject belongs to the control group. The willingness to pay is defined as the unique price list switching point from learning the true state of the world to earning money. We exclude all subjects who show inconsistencies in their behavior, i.e. they switch sides more than once. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.6. Learning about Others - RED type with (mostly) positive Feedback

	(1)	(2)	(3)
Treatment	0.182	0.201	0.436
	(0.294)	(0.299)	(0.345)
Constant	1.983***	2.252***	2.038
	(0.260)	(0.793)	(1.810)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.00313	0.00419	0.126
Observations	124	124	123

Notes: This table reports OLS estimates of subjects' relative learning about another participant's performance on treatment for RED types that received (mostly) positive feedback. Feedback is said to be (mostly) positive when a subject won at least two comparisons. Treatment Dummy equals 1 if a subjects belongs to the treatment and 0 if a subject belongs to the control group. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.7. Learning about Others - BLUE type with (mostly) negative Feedback

	(1)	(2)	(3)
Treatment	-0.112*	-0.110**	-0.0932
	(0.0571)	(0.0549)	(0.0618)
Constant	0.868***	0.565***	0.251
	(0.0516)	(0.112)	(0.438)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.0361	0.116	0.309
Observations	105	105	103

Notes: This table reports OLS estimates of subjects' relative learning about another participant's performance on treatment for BLUE types that received (mostly) negative feedback. Feedback is said to be (mostly) negative when a subject lost all or won only one of the three comparisons. Treatment Dummy equals 1 if a subjects belongs to the treatment and 0 if a subject belongs to the control group. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.8. Willingness to pay: True State of the World

	(1)	(2)	(3)
Treatment	0.0238	0.0219	0.0130
	(0.0362)	(0.0364)	(0.0394)
Constant	0.0625*	0.0168	0.117
	(0.0324)	(0.0787)	(0.347)
IQ Score	No	Yes	Yes
Controls	No	No	Yes
R^2	0.00218	0.00422	0.141
Observations	201	201	199

Notes: This table reports OLS estimates of subjects' willingness to pay to learn the true state of the world on treatment for RED types that received (mostly) positive and BLUE types that received (mostly) negative Feedback. Treatment equals 1 if a subject belongs to the treatment and 0 if a subject belongs to the control group. The willingness to pay is defined as the unique price list switching point from learning the true state of the world to earning money. We exclude all subjects who show inconsistencies in their behavior, i.e. they switch sides more than once. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.9. Different reactions: Bad News vs Good News - Both types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.536***	0.478***	0.468***	0.652***	0.650***	0.681***
	(0.0642)	(0.0664)	(0.0710)	(0.0779)	(0.0794)	(0.0863)
pos. Feedback	-15.25	-15.09	-22.42	-5.199	-6.022	-1.438
	(13.67)	(13.40)	(14.44)	(6.627)	(7.924)	(9.119)
Bayesian Prediction × pos.Feedback	0.341**	0.301*	0.412**	0.117	0.121	0.0933
	(0.162)	(0.160)	(0.171)	(0.114)	(0.117)	(0.129)
Constant	19.75***	3.091	-15.79	16.43***	14.92	-23.14
	(2.833)	(6.665)	(22.80)	(4.858)	(9.295)	(30.89)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.727	0.739	0.787	0.494	0.494	0.596
Observations	165	165	161	165	165	161

Notes: This table reports how subjects' response varied depending on whether they received positive or negative feedback. To study this we add an interaction term (Bayesian Prediction × pos. Feedback). Feedback is said to be positive when a subject won all three comparisons and is said to be negative when a subject lost all three comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.10. Different reactions: Bad News vs Good News - RED types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.0904	0.0730	0.0377	-2.499	-2.227	-1.289
	(0.110)	(0.109)	(0.105)	(2.470)	(2.483)	(2.939)
pos. Feedback	-25.74*	-24.30*	-28.04**	-65.48	-45.81	-15.49
	(14.31)	(14.10)	(13.84)	(40.97)	(45.16)	(54.78)
Bayesian Prediction × pos. Feedback	0.821***	0.737***	0.864***	3.669	3.223	2.118
	(0.192)	(0.194)	(0.193)	(2.517)	(2.552)	(3.029)
Constant	27.42***	9.957	72.34***	50.03*	61.17**	65.01
	(3.949)	(10.10)	(26.02)	(25.91)	(28.05)	(50.36)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.758	0.769	0.885	0.458	0.466	0.641
Observations	80	80	78	80	80	78

Notes: This table reports how subjects' response varied depending on whether they received positive or negative feedback for RED types only. To study this we add an interaction term (Bayesian Prediction × pos. Feedback). Feedback is said to be positive when a subject won all three comparisons and is said to be negative when a subject lost all three comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.11. Different reactions: Bad News vs Good News - BLUE types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.712***	0.665***	0.642***	1.006***	0.873**	1.127**
	(0.0776)	(0.0855)	(0.103)	(0.309)	(0.349)	(0.428)
pos. Feedback	4.311	2.031	-0.138	54.03	31.35	46.86
	(36.61)	(36.52)	(44.11)	(80.87)	(85.53)	(98.05)
Bayesian Prediction × pos. Feedback	0.0148	0.0217	0.0842	-3.670	-2.584	-2.099
	(0.394)	(0.393)	(0.465)	(8.230)	(8.350)	(9.359)
Constant	14.05***	4.404	-56.63*	-10.05	-11.03	-43.02
	(3.667)	(8.397)	(29.62)	(22.94)	(23.02)	(43.51)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.755	0.760	0.823	0.540	0.544	0.665
Observations	85	85	83	85	85	83

Notes: This table reports how subjects' response varied depending on whether they received positive or negative feedback for BLUE types only. To study this we add an interaction term (Bayesian Prediction × pos. Feedback). Feedback is said to be positive when a subject won all three comparisons and is said to be negative when a subject lost all three comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.12. Different reactions: RED vs. BLUE type

	IQ-test performance				World unjust			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bayesian Prediction	0.521***	0.396***	0.386***	0.360***	0.745***	0.673	0.840	0.978
	(0.0966)	(0.0975)	(0.0965)	(0.101)	(0.183)	(0.618)	(0.658)	(0.698)
BLUE	-8.005*	-3.804	-2.926	-2.765	-4.441	-6.315	-2.179	1.421
	(4.350)	(4.314)	(4.279)	(4.433)	(5.726)	(16.43)	(17.35)	(18.45)
Bayesian Prediction × BLUE	0.135**	0.120**	0.101*	0.120*	0.0171	0.0637	-0.0453	-0.137
	(0.0619)	(0.0600)	(0.0597)	(0.0625)	(0.109)	(0.399)	(0.425)	(0.452)
Constant	26.76***	21.70***	7.719	11.87	17.66*	19.75	9.662	39.61
	(6.826)	(6.703)	(8.375)	(19.91)	(9.253)	(19.48)	(23.73)	(36.65)
Feedback	No	Yes	Yes	Yes	No	Yes	Yes	Yes
IQ Score	No	No	Yes	Yes	No	No	Yes	Yes
Controls	No	No	No	Yes	No	No	No	Yes
R2	0.655	0.678	0.686	0.726	0.418	0.418	0.419	0.496
Observations	292	292	292	287	292	292	292	287

Notes: This table reports how subjects' response varied depending on the randomly assigned type. To study this we add an interaction term (Bayesian Prediction × BLUE). Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.13. Different reaction: Initial Overconfidence

	IQ test performance			Unjust world		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Pred.	0.747***	0.663***	0.699***	0.848***	0.848***	0.839***
	(0.0572)	(0.0647)	(0.0690)	(0.0907)	(0.0910)	(0.101)
Over	4.791	7.423	10.24*	4.750	4.499	4.162
	(5.347)	(5.370)	(5.968)	(6.316)	(6.593)	(7.347)
Bayesian P. × Over	-0.168**	-0.145*	-0.180**	-0.130	-0.131	-0.114
	(0.0783)	(0.0778)	(0.0841)	(0.119)	(0.119)	(0.130)
Constant	14.43***	-7.594	-1.936	7.227	8.630	13.74
	(4.315)	(9.377)	(26.83)	(4.693)	(11.29)	(33.24)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.615	0.627	0.688	0.445	0.445	0.509
Observations	223	223	219	223	223	219

Notes: This table reports how the subjects' response varied depending on initial overconfidence. *Overconfident* is a dummy equal to one if a subject had an expected rank at least 1.5 ranks below (i.e. better) than the actual rank. The dummy is equal to zero if the expected rank is accurate, i.e. the expected rank is neither more than 1.5 ranks better nor 1.5 ranks worse than the actual rank. To study differences in reactions, we add an interaction term (Bayesian Prediction × over). Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.14. The effect of initial overconfidence on unjust world belief

	RED type		BLUE type	
	(1)	(2)	(3)	(4)
Overconfident	-11.12	-14.96*	-0.783	3.514
	(7.307)	(8.955)	(6.777)	(8.562)
Constant	-4.219	13.92	110.7***	122.2**
	(17.19)	(45.50)	(20.77)	(50.21)
Rank & Feedback	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
R^2	0.414	0.528	0.544	0.638
Observations	120	119	103	100

Notes: This table reports the effects of initial overconfident beliefs on the unjust World belief. *Overconfident* is a dummy, that is equal to one if a subject had an expected rank 1.5 ranks above the actual rank. The dummy is equal to zero if the expected rank is accurate, i.e. the expected rank is neither more than 1.5 ranks better nor 1.5 ranks worse than the actual rank. Standard errors in parentheses. Rank & feedback controls encompass variables for the received feedback and the actual rank of the participants. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.15. Effects (Mostly) Bad vs. (Mostly) Good News - Both Types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.522***	0.473***	0.482***	0.717***	0.712***	0.702***
	(0.0452)	(0.0473)	(0.0504)	(0.0716)	(0.0720)	(0.0755)
(mostly) pos. Feedback	-9.351	-9.450	-8.223	-4.974	-6.574	-5.865
	(6.519)	(6.421)	(6.605)	(5.717)	(6.104)	(6.494)
Bayesian Prediction × (mostly) pos. Feedback	0.285***	0.268***	0.266***	0.114	0.124	0.132
	(0.0856)	(0.0845)	(0.0870)	(0.109)	(0.110)	(0.116)
Constant	20.11***	6.047	10.15	13.56***	8.639	41.79
	(2.348)	(5.048)	(18.17)	(4.022)	(7.674)	(26.28)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.686	0.697	0.731	0.416	0.417	0.493
Observations	292	292	287	292	292	287

Notes: This table reports how subjects' response varied depending on whether they received (mostly) positive or (mostly) negative feedback. To study this we add an interaction term (Bayesian Prediction × (mostly) pos. Feedback). Feedback is said to be (mostly) positive when a subject won at least 2 comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.16. Effects (Mostly) Bad vs. (Mostly) Good News - RED Types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.270***	0.250***	0.273***	0.483	0.550*	0.575*
	(0.0672)	(0.0676)	(0.0712)	(0.293)	(0.304)	(0.344)
(Mostly) pos. feedback	-14.11*	-13.78*	-13.05*	4.081	6.449	-9.455
	(7.501)	(7.448)	(7.557)	(14.15)	(14.42)	(15.87)
Bayesian Prediction × (mostly) pos. feedback	0.516***	0.486***	0.498***	0.149	0.0981	0.336
	(0.107)	(0.108)	(0.111)	(0.360)	(0.365)	(0.412)
Constant	26.49***	14.17*	42.77**	17.55**	25.14**	32.48
	(3.301)	(7.646)	(21.42)	(6.767)	(11.09)	(33.30)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.683	0.690	0.766	0.362	0.365	0.492
Observations	150	150	148	150	150	148

Notes: This table reports how subjects' response varied depending on whether they received (mostly) positive or (mostly) negative feedback for RED types only. To study this we add an interaction term (Bayesian Prediction × (mostly) pos. Feedback). Feedback is said to be (mostly) positive when a subject won at least 2 comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 3.F.17. Effects (Mostly) Bad vs. (Mostly) Good News - BLUE Types

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	0.710***	0.660***	0.662***	0.644***	0.619***	0.609***
	(0.0569)	(0.0640)	(0.0728)	(0.198)	(0.196)	(0.213)
(Mostly) pos. feedback	-9.776	-9.069	-13.70	-5.567	-12.92	-7.874
	(14.49)	(14.40)	(15.75)	(14.78)	(15.06)	(16.01)
Bayesian Prediction × (mostly) pos. feedback	0.180	0.165	0.215	-0.165	-0.0789	-0.186
	(0.169)	(0.168)	(0.184)	(0.302)	(0.302)	(0.328)
Constant	13.37***	4.134	-20.22	19.08	1.849	11.92
	(3.048)	(6.302)	(21.83)	(13.68)	(15.97)	(34.23)
IQ Score	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.749	0.754	0.793	0.484	0.499	0.593
Observations	142	142	139	142	142	139

Notes: This table reports how subjects' response varied depending on whether they received (mostly) positive or (mostly) negative feedback for BLUE types only. To study this we add an interaction term (Bayesian Prediction × (mostly) pos. Feedback). Feedback is said to be (mostly) positive when a subject won at least 2 comparisons. Columns (1) to (3) report results on the IQ-test performance belief and Columns (4) to (6) on the unjust world belief. Standard errors in parentheses. Controls include variables for age, gender, education, field of study, income, BIG-5 personality traits, and the Narcissism Score. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 3.G Instructions

Note: Translated by the authors. We first record the instructions for the treatment group and subsequently show where the instructions differed for subjects in the control group.

Treatment

Day 1: Intelligence Test and Surveys

Page 1 - Welcome:
Welcome to the experiment!
The experiment consists of two parts:

- Part 1 takes place today
- Part 2 takes place tomorrow, [Date]. The second part will take place in the BonnEconLab. Please arrive at the lab 15 minutes before the experiment starts.

From the time that you start this part of the experiment, you will need 2 hours to complete it. You have to finish the study in one take, meaning discontinuities are not allowed. To finish Part 1 without problems, you should ensure a stable internet connection. Please do not execute this experiment on a smartphone.
You are only allowed to participate in the experiment tomorrow if you finish this part.

Page 2 - Payment:
You will receive a fixed payment of X EUR upon completion of both parts of the experiment. Depending on your decisions you can earn additional money. We will explain when and how you can earn additional money at the relevant stages of the experiment. Again, you must complete the entire experiment in order to receive any money.
Your total payment (the fixed payment plus the additional payments) will be given to you after completing Part 2 in the BonnEconLab.

Page 3:
As soon as you start Part 1, you have to complete it without a break. You are only allowed to participate in the experiment tomorrow if you finish this part. Please do not execute this experiment on your smartphone.

If you have any questions, please feel free to send an email to: exp_2019@uni-bonn.de

If you are ready to start the experiment, press NEXT.

Page 4 - Demographics:
Please answer the following questions:

•How old are you?
•Which gender do you identify with?

–Male
–Female
–Other
–Prefer Not to say

•Is German your mother tongue?

–Yes
–No

•What is your highest educational qualification?

–Without school-leaving qualification
–Lower secondary education
–Secondary school certificate
–A-Levels
–University Degree (Bachelor/Master/Diploma)
–PhD
–Different certificate
–Prefer not to say

•If you study, what category describes your subject of study best?

–I did not study
–Law
–Economics / Business
–Natural Sciences
–Engineering, Maths, Informatics
–Social Sciences
–Music, Art
–Languages or Cultural studies
–Media, Communication

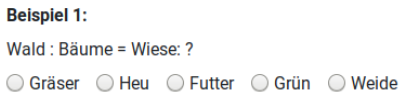
–Other

- What is your monthly household net-income?
- On a scale form 0 to 10, how willing are you to take risks? 0 means that you are not willing to take risks at all and 10 means that you are more than willing to do so.

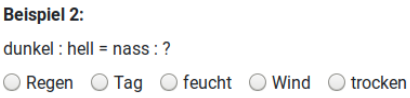
Page 5 - Transition:
Thank you for answering. Next you will complete a test that is introduced on the following pages.

Page 6 - Introduction Test:
The test consists of three parts. Each part has 20 exercises and a time limit. If you reach the time limit, the part will automatically come to an end. For each correct answer you receive 10 cents. In total you can earn up to 6 EUR.
It is likely that you will not be able to answer all questions within the time limit. You should not be concerned about this.

Page 7 - Test Part 1:
Part 1 consists of 20 similar exercises. In each exercise you will see 3 words. The first and second words are related in some way. Your job is to find the word whose context corresponds to the third word in the way that the first and second words corresponded.
Please look at the two examples to get a better understanding of the exercise:



The relation between Wald (Forest) and BÃ€ume (Trees) is that there are many trees in a forest. From the suggested options you now have to find the word that is similarly related to the third word Wiese (meadow). The correct answer is GrÃ€ser (Grass).



Dunkel (Dark) is the opposite of hell (bright), so you have to find the opposite of nass (wet). The right answer to example 2 is trocken (dry).

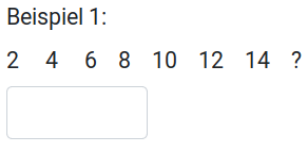
In what follows there are 20 of above described exercises. For each correct answer you receive 10 cents. You have 5 minutes to answer as many of the 20 exercises as possible. After the time is up you will be automatically forwarded to the next part. You can use the Back button to review and adjust your answer to a previous exercise.

Part 1 of the test begins as soon as you click NEXT.

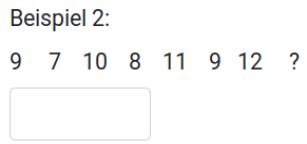
Page 8 - Test Part 1:
Participants had 5 mins to work on the 20 exercises.

Page 9 - End Test Part 1:
Your time is up. Part 1 of the test is over.

Page 10 - Test Part 2:
The second part of the experiment consists of 20 exercises of the same type. We will show you sequences of integers. The sequences follow a rule and each sequence can be extended using this rule. Your exercise will be to find the next integer in the sequence.
Please look at the following two examples to get a better understanding of the exercise:



In this sequence of integers each number is is greater than the one before by 2: 4 is greater than 2 by 2, 6 is greater than 4 by 2, and so on. The solution to this exercise is 16.



In this sequence of integer you have to alternate between subtracting 2 and adding 3: $9-2=7$; $7+3=10$; $10-2=8$; $8+3=11$; $11-2=9$; $9+3=12$; $12-2=10$. Thus, the right answer is 10.

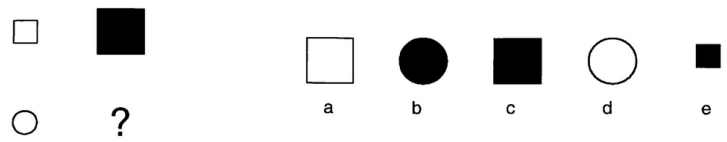
In what follows there are 20 of above described exercises. For each correct answer you receive 10 cents. You have 7 minutes to answer as many of the 20 exercises as possible. After the time is up you will be automatically forwarded to the next part. You can use the Back button to review and adjust your answer to a previous exercise.
Part 2 of the test begins as soon as you click NEXT.

Page 11 - Test Part 2:
Participants had 7 mins to work on the 20 exercises.

Page 12 - End Test Part 2:
Your time is up. Part 2 of the test is over.

Page 13 -Test Part 3:
As in the first two parts, the last part also consists of 20 exercises. On the left side you will see a sequence of figures. The sequences are built Using a certain rule. On the right side you will see five other figures. Out of these five addiditional figures you have to find the one that should replace the question mark on the left side, i.e. that fits in with the sequence.
Please look at the two following examples to get a better understanding of the exercise:

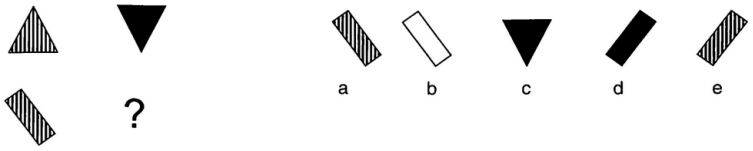
Beispiel 1:



○a ○b ○c ○d ○e

Focusing on the first row we observe that the small white square changes to a big black square. The form stays the same but the color and the size change. Following this logic, the small white circle should change into a big black circle. Hence, the correct answer is choice b.

Beispiel 2:



○a ○b ○c ○d ○e

In this example the triangle in the first row is mirrored (turned upside down) and blackened. Thus, the rectangle in the second row has to be mirrored and blackened as well. This is done in solution d, which is therefore the correct answer.
In what follows there are 20 of above described exercises. For each correct answer you receive 10 cents. You have 7 minutes to answer as many of the 20 exercises as possible. After the time is up you will be automatically forwarded to the next part. You can use the Back button to review and adjust your answer to a previous exercise.

Part 3 of the test begins as soon as you click NEXT.

Page 14 - End Test:
Your time is up. You completed all three parts of the test.
To finish todays part of the experiment, please fill out the following questionnaires.
Page 15 - BIG 5:
Below we list different characteristics a person can have. Some of these characteristics probably apply to your personality while others will not apply at all. In what follows we ask you to state how much the characteristics apply to you. Please give your answer on a scale from 0 = Does not apply at all to 7 = fully applies.

See: BIG 5 - 20 item questionnaire Topolewska-Siedzik, Skimina, Strus, Cieciuch, and Rowiński (2014).

Page 15 - Narcissism:
Please answer the degree to which the following statements apply to you. You must give your answers on a scale from 0 = does not apply to 5 = fully applies.

See: Narcissistic Personality Inventory - 16 item questionnaire Ames, Rose, and Anderson (2006).

Page 16 - End Day 1:
Thank you! You finished the first part of the experiment. You will receive your payment at the end of tomorrow's experiment. Please arrive at the BonnEconLab 15 minutes before the start of the experiment.

Day 2: Feedback, Posterior and Consequences

Page 17 - Welcome Day 2:
Welcome back.
It is forbidden to talk to other people during the experiment. Please turn off your phones. If you have any questions during the experiment, please hold out you hand. One of the supervisors will come to you and answer your question.
Reminder: This experiments consists of two parts.

- You completed Part 1 yesterday
- Part 2 takes place right now

Click NEXT to start the experiment.

Page 18 - Payment:
Similar to yesterday you can earn additional money during the second part of the

experiment. The total payment will be given to you at the end of the experiment. How much additional money you earn is going to depend on your decisions. During the experiment you will face several payment relevant decision. How you can earn additional money will be explained at the relevant stages in more detail.
To earn as much money as possible, it is important that you read the instructions carefully.

Page 19 - Estimates:
In some parts of the experiment, we ask you to estimate how likely certain statements are. More specifically, we will ask you to state what you think the probability is that a certain statement is true. The exact statements will be described to you at the relevant stages. It is important to know that your estimates are relevant for your payment. For each estimate in the experiment, you can earn up to 2 EUR. The exact formula is:

$$Earnings \;=\; 2 - 2(I(true) - \frac{belief}{100})^2$$

Even if this formula looks complicated, it is always true that:
The closer your estimate is to the true value, the more money you will earn.

Page 20 - IQ-test:
Success in life depends on many factors. One very important one is intelligence. Many studies show that intelligence plays an important role for a successful life: intelligent people receive better school leaving certificates, have more professional success and earn more. Thus, intelligence is a driving factor for a successful life.

IQ-test
The test you completed yesterday is part of a widely used IQ-test. The parts you completed measured three different types of intelligence. Part 1, in which you had to find the relationships between pairs of words, measured your verbal intelligence. Part 2, in which you had to complete sequences of integers, measured your numerical intelligence. The third part, in which you completed sequences of figures, measured your figural-spatial intelligence. In contrast to many other intelligence tests, our IQ-test takes several facets of intelligence into account.

Comparison Group
We randomly assigned nine other participants to you. These nine other participants completed the same IQ-test within a different experiment. For all 10 participants (you plus the nine others) in your group we calculated the point score of the IQ-test, where each right answer is one point. Based on this score, we ranked all group members. The participant with the highest point score is ranked as number one. The participant with the second highest point score is ranked as number two and so

on. In the unlikely case that two or more participants have the identical IQ-Score, a computer randomly determines who gets the highest rank.

Page 20 - Prior IQ-test performance:
How do you think you performed compared to the other participants?

We will ask you to make several estimations. As explained before, you can earn additional money for your stated belief. At the end of the experiment, a computer will randomly choose and pay out one of the following estimations. You can earn up to 2 EUR. The closer your estimation is to the true value the more money you can earn. Thus, the probability you state should be as correct as possible.

Page 21 - Prior IQ-test performance:
What do you think is the likelihood that your IQ-test Score ranked in the upper half of the group? In other words, please state the probability that you ranked number one, two, three, four or five?

Answer: XXX %

Page 22 - Prior IQ-test performance:
You stated that you are ranked in the upper half of the IQ-Ranking with a probability of XXX %. Now, we ask you to distribute the probability among the five upper ranks.
What is the likelihood that you are ranked as...
Number 1: a %
Number 2: b %
Number 3: c %
Number 4: d %
Number 5: e %

Page 23 - Prior IQ-test performance:
You stated that you are ranked in the lower half of the IQ-Ranking with a probability of 100 - XXX %. Now, we ask you to distribute the probability among the five lower ranks.
What is the likelihood that you are ranked as...
Number 6: a %
Number 7: b %
Number 8: c %
Number 9: d %
Number 10: e %

Page 24 - End Prior IQ-test performance:
Thank you for your estimations.

In what follows you will receive information about the further procedure of the experiment. Please read the instructions carefully. It is essential that you fully understand them. If you have any questions, please do not hesitate to ask.

Page 25 - Instructions Feedback:
Feedback: The comparison
You will receive feedback about how you performed in the IQ-test compared to others in your group. To be more precise, we will make three comparisons between you and three randomly picked people from your group. You will either receive positive or negative feedback. Whether you receive positive or negative feedback depends on three factors:

- Your and the other person's point score in the IQ-test
- Your and the other persons type
- The world in which you and all other group members live

On the next pages, we will explain each factor in more detail.

Page 26 - Instructions Feedback:
Score on the IQ-test.

The point score on the IQ-test plays a central role for the comparisons. The basic idea is that you will receive positive feedback if you were better than the other person and negative feedback if you had a lower point score in the IQ-test. But, this is not always the case:
(If participant RED type:)
There is the possibility that you win the comparison although you have a lower point score in the IQ-test than the person you are compared with.
(If participant BLUE type:)
There is the possibility that you lose the comparison although you have a higher point score in the IQ-test than the person you are compared with.
(Both again:)
Under what circumstances this happens will be explained in the following pages.

Page 25 - Instructions Feedback:
Types
Every participant is either a RED or BLUE type. Both types are equally represented in the group, i.e. 5 participants in your group are RED and 5 are BLUE. We will tell you what type you are. Your type stays the same for the rest of the experiment.

You are a RED/BLUE type.

RED and BLUE types are not the same. There exists the possibility that RED types are privileged over BLUE types. On the next page you will learn when this is the case.

Page 26 - Instructions Feedback:
World

In this experiment there exist two types of worlds in which you theoretically can live in: an unjust and a just world.
At the beginning of the experiment, one of the two worlds was randomly chosen. This means that the probability to be in the unjust world is 50% and the probability to be in the just world is also 50%. You will stay in the randomly chosen world for the rest of the experiment. The two worlds differ:
In the unjust world RED types are privileged and BLUE types are discriminated. This means

- If a RED type is compared with a BLUE type, the RED type always wins - independent of the point score in the IQ-test.
- If two persons of the same type are compared, the person with the higher point score wins.

In the just world both types are equal, i.e. there exists no discrimination.

- The person with the higher IQ-test score always wins.

Importantly: You will never know with whom you were compared. In particular, you will never learn the type of the other person. You will also not learn in which world you live. You will only receive feedback about whether you won or lost the comparison.

Page 27 - Instructions Feedback:
Recap:
In a few moments you will receive feedback about your intelligence. To do so, you will be compared three times with three random people of your group of ten. You will not learn if the other person is of RED or BLUE type. Further, you will not fully learn in which world you live.
(If participant RED type:)
Potential reasons for winning a comparison are:

- You had a higher point score in the IQ-test.
- You live in the unjust world and the other person was a BLUE type.

Potential reasons for losing a comparison are:

- You had a lower point score in the IQ-test.

(If participant BLUE type:)
Potential reasons for losing a comparison are:

- You had a lower point score in the IQ-test.
- You live in the unjust world and the other person was a RED type.

Potential reasons for winning a comparison are:

- You had a higher point score in the IQ-test.

On the next page we will ask you some control questions. If you think that you fully understood the instruction please press NEXT.

Page 28 - Control Questions:

- What is your type?
- How many RED and how many BLUE types are in your group of ten?

–2 RED & 8 BLUE
–5 RED & 5 BLUE
–8 RED & 2 BLUE

- Are you privileged or discriminated in the unjust world?

(If participant RED type:)

- Assume that you will be compared with a person who has a higher point score in the IQ-test. In which world will you for certain lose the comparison?

(If participant BLUE type:)

- Assume that you will be compared with a person who has a lower point score in the IQ-test. In which world will you for certain win the comparison?

Page 29 - Feedback:
On the next page you will receive your feedback.
Page 30 - Feedback:
Comparison 1:
You won/lost the comparison
Comparison 2:
You won/lost the comparison
Comparison 3:
You won/lost the comparison

Page 31 - Repeat Feedback:
How many comparisons did you win?

How many comparisons did you lose?

Page 31 - Posterior:
After receiving your feedback, what do you think:

- How well did you perform in the IQ-test?
- In which world are you living?

One of the following two estimations will be randomly chosen and paid out at the end of the experiment.
You can earn up to 2 EUR. The formula that determines your payment is the same as before.

IQ-test performance
What do you think is the likelihood that you IQ-test Score ranked in the upper half of the group? In other words, please state the probability that you ranked number one, two, three, four or five?

Answer: XXX %

Unjust world
What do you think is the likelihood that you are living in the unjust World? In other words, please state the probability that you potentially received distorted feedback?
Answer: YYY %

Page 31 - Posterior:
Thank you. In the next step, we ask you to participate in a game. You can earn additional money. We will explain the rules of the game on the next page.

Page 32 - Effort task:
You can earn additional money in this exercise. The task is simple and does not require any special skills. More specifically, intelligence does not play a role in the exercise.

Task
You will have up to 5 minutes to pull as many sliders as possible to the number 500. To do this you can either use your computer mouse or the arrow keys on your keyboard. Please pull the following slider to 500 to get a better understanding of the task:

[Example Slider]

Comparison group
In an early experiment other people did the same task as you are about to do. We will randomly draw 9 people from this group. Your performance in the slider task will be compared with one of these nine people.

Important: You are still type RED/BLUE

Comparison
As before, the comparison depends not only on your performance but also on your type and the world that you are living in:

- In the just world only your performance matters. If you pulled more sliders to 500 than your partner, you win. If you pulled less sliders to 500, you lose. In the unlikely case of a draw a computer randomly decided whether you win or lose.
- In the unjust world the type of the other person is of importance:
  - (If participant is RED type): If the other person is BLUE type, you always win. If the person is RED, the one with more sliders pulled to 500 will win.
  - (If participant is BLUE type): If the other person is RED type, you always lose. If the other person is BLUE, the one with more sliders pulled to 500 will win.

Payment
If you win the comparison you earn 4 additional EURs.

It is your decision for how long and how many sliders you try to pull to 500.

Page 32 - Effort task:
Page with 86 Sliders.

Page 33 - Effort task:
The slider task is over. At the end of the experiment you will learn whether you earned the additional 4 EUR or not. In a next step you will observe the feedback of a different person.

Page 34 - Social learning:
We will now show you the feedback that a different person received. This person did the same experiment, meaning they person completed the identical IQ-test and received feedback about their performance.
Information about the different person:

- This person is in a different group of ten
- This person lives in the same world

- Reminder: After you received feedback, you said that you live in the unjust world with a likelihood of XX%
- This person is BLUE type

On the next page we will show you the other person's feedback. Afterwards we ask you to make two estimations: one about the intelligence of the other person and the other about the world you both live in.

Page 35 - Social learning:
(If participant RED type):
Comparison 1: Different person lost comparison.
Comparison 2: Different person lost comparison.
Comparison 3: Different person lost comparison.
(If participant BLUE type):
Comparison 1: Different person won comparison.
Comparison 2: Different person won comparison.
Comparison 3: Different person won comparison.

Page 36 - Social learning:
After you observed the feedback of the other person, what do you think:

- How well did the other person perform on the IQ-test?
- In which world are you and the other person living?

One of the following two estimations will be randomly chosen and paid out at the end of the experiment.
You can earn up to 2 EUR. The formula that determines your payment is the same as before.

IQ-test performance
What do you think is the likelihood that the other person is ranked in the upper half of her group? In other words, please state the probability that the other person is ranked number one, two, three, four or five?

Answer: XXX %

Unjust World
What do you think is the likelihood that you and the other person are living in the unjust World? In other words, please state the probability that you and the other person potentially received distorted feedback?
Answer: YYY %

Page 37 - Willingness to pay:
Thank you for your estimations.
Now we give you the possibility to learn in which world you live in during the experiment. This means you can learn whether you lived in the just world and received true, undistorted feedback about your IQ-test performance or you lived in the unjust world and received distorted feedback.

Page 38 - Willingness to pay:
On the next page you have to make 21 decisions. In each decision you will have to choose between two options. One option for all 21 decisions stays the same while the other varies. The constant option is that you learn in which world you lived during the experiment. The other option is a monetary value that you either receive or have to pay. At the end of the experiment we will randomly choose one of the 21 decisions and implement your choice.

Example [Option 1: Learn World; Option 2: Pay 10 Cent]

In the example you have to decide between paying 10 cents and learning the state of the world.

If you are ready click NEXT.

Page 39 - Willingness to pay:
Price list:
Decision 1: [Option 1: Learn World; Option 2: receive 50 CENT]
...
Decision 21: [Option 1: Learn World; Option 2: pay 1.50 EUR]

Page 40 - Questionnaire:
Please answer the following questions to wrap up the experiment:

- What best describes your sexual orientation?
  - Heterosexual
  - Bisexual
  - Homosexual
  - Asexual
  - Other
  - Prefer not to say
- Were you born in Germany?
- Are both your parents born in Germany?

•Are you religious? (0=Not at all, 7= very)
•What is your religious denomination?

–Christianity
–Islam
–Buddhist
–Jewish
–Hindu
–Different denomination
–Without denomination
–No response

•Did you grow up in an urban or rural area? (0 = big city, 6 = small village)
•What is your father's highest school-leaving certificate?

–Without school-leaving qualification
–Lower secondary education
–Secondary school certificate
–A-Levels
–University Degree (Bachelor/Master/Diploma)
–PhD
–Different certificate
–Prefer not to say

•What is your mother's highest school-leaving certificate?

–Without school-leaving qualification
–Lower secondary education
–Secondary school certificate
–A-Levels
–University Degree (Bachelor/Master/Diploma)
–PhD
–Different certificate
–Prefer not to say

•Compared to the average German household, how would you describe your parents' household income? For your information, the average gross household income in Germany is 4200 EUR per month.

–Much lower

–Lower
–About the same
–More
–Much More
–I don't know
–Prefer not to say

Control

Participants in the control sessions were randomly matched with one of the 292 participants in the Treatment. This means the below introduced Person Z was a participant from the Treatment. When all control sessions are finished we will have a one to one matching between participants in the ego-relevant Treatment and the control in which participants observe and make estimations about an unknown other person.
Most of the experiment stayed the same for the participants in the control group. Therefore, we only present the instructions for the one page on which the other person was introduced.

Page 25 - Instructions Feedback:
Person Z

So far the experiment was about your intelligence. This no longer is the case. The rest of the experiment is about a randomly chosen other person. We will call this person Person Z from now on.

Person Z already completed the experiment. Person Z did the same IQ-test as you. Furthermore, Person Z is part of a different group of ten for which we calculated an IQ-ranking based on the performance in the IQ-test. As you have no further information, you also do not know how many points Person Z scored in the IQ-test.

Summary:
The rest of the experiment is concerned with Person Z. Person Z ..

- was randomly allocated to you,
- did an identical IQ-test,
- is part of a different group of 10,
- and you have no information about Person Z's performance on the IQ-test.

In the following you will observe feedback about Person Z's intelligence. We will explain the procedure in more detail on the following pages.

The remaining instructions followed the same logic as above, with the only difference being that whenever we talked about 'you' in the treatment, we replaced it with Person Z in the control.

References

Alesina, Alberto, and George-Marios Angeletos. 2005. "Fairness and redistribution." *American economic review* 95 (4): 960–980. [113, 117]

Alesina, Alberto, and Eliana La Ferrara. 2005. "Preferences for redistribution in the land of opportunities." *Journal of public Economics* 89 (5-6): 897–931. [113, 117]

Ames, Daniel R, Paul Rose, and Cameron P Anderson. 2006. "The NPI-16 as a short measure of narcissism." *Journal of research in personality* 40 (4): 440–450. [174]

Arkin, Robert, Harris Cooper, and Thomas Kolditz. 1980. "A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations 1." *Journal of Personality* 48 (4): 435–448. [116]

Barron, Kai. 2016. "Belief updating: Does the'good-news, bad-news' asymmetry extend to purely financial domains?" [116]

Benabou, Roland, and Jean Tirole. 2006. "Belief in a just world and redistributive politics." *Quarterly journal of economics* 121 (2): 699–746. [113, 117]

Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." *Quarterly Journal of Economics* 117 (3): 871–915. [116]

Billett, Matthew T, and Yiming Qian. 2008. "Are overconfident CEOs born or made? Evidence of self-attribution bias from frequent acquirers." *Management Science* 54 (6): 1037–1051. [116]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg registration and organization online tool." *European Economic Review* 71: 117–120. [122]

Brunnermeier, Markus K, and Jonathan A Parker. 2005. "Optimal expectations." *American Economic Review* 95 (4): 1092–1118. [116]

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini. 2013. "Overconfidence and social signalling." *Review of Economic Studies* 80 (3): 949–983. [116]

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9: 88–97. [122]

Chen, Zhuoqiong Charlie, and Tobias Gesche. 2017. "Persistent bias in advice-giving." *University of Zurich, Department of Economics, Working Paper*, (228): [116]

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics* 22 (2): 369–395. [116, 124]

Coutts, Alexander, Leonie Gerhards, and Zahra Murad. 2019. "No one to blame: Self-attribution bias in updating with two-dimensional uncertainty." [116]

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* 33 (1): 67–80. [116]

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. "Investor psychology and security market under-and overreactions." *the Journal of Finance* 53 (6): 1839–1885. [116]

Di Tella, Rafael and Perez-Truglia, Ricardo and Babino, Andres and Sigman, Mariano. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review* 105 (11): 3416–42. [116]

Doukas, John A, and Dimitris Petmezas. 2007. "Acquisitions, overconfident managers and self-attribution bias." *European Financial Management* 13 (3): 531–577. [116]

Eil, David, and Justin M Rao. 2011. "The good news-bad news effect: asymmetric processing of objective information about yourself." *American Economic Journal: Microeconomics* 3 (2): 114–38. [114, 116, 118, 120, 127, 128, 144]

Exley, Christine L. 2016. "Excusing selfishness in charitable giving: The role of risk." *Review of Economic Studies* 83 (2): 587–628. [116]

Exley, Christine L. 2020. "Using charity performance metrics as an excuse not to give." *Management Science* 66 (2): 553–563. [116]

Exley, Christine L, and Judd B Kessler. 2019. "Motivated errors." Working paper. National Bureau of Economic Research. [116]

Falk, Armin, Thomas Neuber, and Nora Szech. 2020. "Diffusion of Being Pivotal and Immoral Outcomes." *Review of Economic Studies*, [116]

Gervais, Simon, and Terrance Odean. 2001. "Learning to be overconfident." *Review of Financial Studies* 14 (1): 1–27. [116]

Gill, David, and Victoria Prowse. 2012. "A structural analysis of disappointment aversion in a real effort competition." *American Economic Review* 102 (1): 469–503. [121]

Gneezy, Uri, Elizabeth A Keenan, and Ayelet Gneezy. 2014. "Avoiding overhead aversion in charity." *Science* 346 (6209): 632–635. [116]

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the self." *Games and Economic Behavior* 120: 311–324. [116]

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information avoidance." *Journal of Economic Literature* 55 (1): 96–135. [116]

Haisley, Emily C, and Roberto A Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and Economic Behavior* 68 (2): 614–625. [116]

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2018. "Unrealistic expectations and misguided learning." *Econometrica* 86 (4): 1159–1214. [148]

Hestermann, Nina, and Yves Le Yaouanq. 2020. "Experimentation with Self-Serving Attribution Biases." *American Economic Journal: Microeconomics*, [148]

Hilary, Gilles, and Lior Menzly. 2006. "Does past success lead analysts to become overconfident?" *Management science* 52 (4): 489–500. [116]

Hoffmann, Arvid OI, and Thomas Post. 2014. "Self-attribution bias in consumer financial decision-making: How investment returns affect individuals' belief in skill." *Journal of Behavioral and Experimental Economics* 52: 23–28. [116]

Kim, Y Han Andy. 2013. "Self attribution bias of the CEO: Evidence from CEO interviews on CNBC." *Journal of Banking & Finance* 37 (7): 2472–2489. [116]

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American economic review* 90 (4): 1072–1091. [116]

Köszegi, Botond. 2006. "Ego utility, overconfidence, and task choice." *Journal of the European Economic Association* 4 (4): 673–707. [116]

Lerner, Melvin J. 1980. "The belief in a just world." In *The Belief in a just World*. Springer, 9–30. [117]

Li, Feng. 2010. "Managers' self-serving attribution bias and corporate financial policies." *Available at SSRN 1639005*, [116]

Libby, Robert, and Kristina Rennekamp. 2012. "Self-serving attribution bias, overconfidence, and the issuance of management forecasts." *Journal of Accounting Research* 50 (1): 197–231. [116]

Mezulis, Amy H, Lyn Y Abramson, Janet S Hyde, and Benjamin L Hankin. 2004. "Is there a universal positivity bias in attributions? A meta-analytic review of individual, devel-

opmental, and cultural differences in the self-serving attributional bias." *Psychological bulletin* 130 (5): 711. [116]

Miller, Dale T, and Michael Ross. 1975. "Self-serving biases in the attribution of causality: Fact or fiction?" *Psychological bulletin* 82 (2): 213. [116]

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." Working paper. National Bureau of Economic Research. [116]

Sandel, Michael J. 2020. *The tyranny of merit: What's become of the common good?* Penguin UK. [113]

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and self-deception." *Nature human behaviour* 3 (10): 1055–1061. [116]

Sharot, Tali, Christoph W Korn, and Raymond J Dolan. 2011. "How unrealistic optimism is maintained in the face of reality." *Nature neuroscience* 14 (11): 1475. [116]

Topolewska-Siedzik, Ewa, Ewa Skimina, Włodzimierz Strus, Jan Cieciuch, and Tomasz Rowiński. 2014. "The short IPIP-BFM-20 questionnaire for measuring the big five." *Annals of Psychology* 17 (01): 385–402. [174]

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." *American Economic Review* 110 (2): 337–61. [114, 116]

Zuckerman, Miron. 1979. "Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory." *Journal of personality* 47 (2): 245–287. [116]