Essays on Beliefs

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

Lasse Simon Stötzer

aus Hannover

Bonn, 2020

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Lorenz Goette
Zweitreferent:	Prof. Dr. Armin Falk
Tag der mündlichen Prüfung:	27. November 2020

Acknowledgments

I would like to thank my advisors Lorenz Goette and Armin Falk for their guidance and trust. Our discussions helped me to get an idea of what constitutes 'good' research and motivated me to always dig deeper. I would also like to thank Florian Zimmermann for his suggestions and for creating an environment in which I felt comfortable and respected.
I benefited greatly from the environment at the Bonn Graduate School of Economics and the support from briq - Institute on Behavior & Inequality, ECONtribute: Markets & Public Policy, and the Institute for Applied Microeconomics. In particular, I want to thank Stefanie Sauter, Markus Antony, Andrea Reykers, and Silke Kinzig who helped me on various administrative matters.
I owe special thanks to Sven Heuser. Thanks for being a great friend, co-author, and fellow Panini Sticker enthusiast. Doing a Ph.D. can be frustrating at times, which makes it crucial to have friends who distract and support you. Thanks to Renate Boden, Laura Ehrmanntraut, Anita Gohdes, Meret Hesse, Gerrit Jung, Carla Kaspari, Marius Kulms, Karolin Kupfer, Anna Lane, Marwa Mahran, Tim Maurer, Signe Moe, Marcel Mucha, Nazli Oral, Julia Scheibe, Paula Schulze Brock, Renske Stans, Kathrin Stegherr, Christopher Szwabczynski, and Moritz von Stetten for doing exactly this.
Finally, I want to thank my family - Ursel, Friedel, and Lukas - for their love, trust, and encouragement.

Contents

List of Figures

List of Tables

Introduction

Beliefs govern how individuals interact with the world around them. Their pivotal role in decision-making processes has led to great interest in how they are formed and arguably of more importance, how they are changed. The conceptual framework of the neoclassical model relies on the assumption that individuals form their beliefs about an unobservable state of the world by incorporating all available information and that they process this information according to a normative updating rule. Together with the assumption that preferences are stable and egoistic, the resulting theoretical framework is in many cases a powerful predictor of human behavior. However, the simplicity of the model often does not live up to the complexity of human decision making and therefore produces predictions that are at odds with reality.

The field of behavioral economics has sought to add depth to the neoclassical model by adapting the notions of maximization and preferences to incorporate systematic anomalies observed in the real world. A large literature within the field has also concerned itself with belief formation and persistence. Building on this literature, I endeavor to advance our understanding of human decision-making by studying the role of subjective beliefs. In particular, this thesis studies how biased beliefs can be changed and how they are formed. In chapter 1, I examine whether confronting people with contrasting viewpoints can change their beliefs about the people who hold those viewpoints. This question has grown in relevance as polarized beliefs, ensuing from a society in which individuals seek out information that predominantly confirms their beliefs, have risen to be a defining phenomenon of recent years. In chapters 2 and 3, I study how motives mold individuals' subjective beliefs. If individuals are motivated to protect certain beliefs, the normative updating rule of the neoclassical model is insufficient. To address these scenarios, this thesis utilizes multiple experimental methods and draws insights from the fields of psychology, political science, and sociology.

Chapter 1: "Now we're talking: The impact of political face-to-face conversations" is motivated by the popular narrative that political polarization increased over the last couple of years in many Western democracies. Sunstein (2018) and Bishop (2009) argue that the source of rising polarization is the sorting of individuals into environments in which they mostly interact with like-minded individuals. If

the lack of social interaction with contrary-minded individuals causes polarization and increasing animosity between political camps, enhancing opportunities for communication should help to counteract this trend.
In joint work with Sven Heuser, we address whether a political face-to-face conversation helps reduce stereotypes about people with different political views, bolsters perceived social cohesion and changes political attitudes. Leveraging a large scale intervention in Germany with over 19,000 registered participants, we present causal evidence on these questions. We find that the effect heavily depends on whether like- or contrary-minded pairs met. A conversation between contrary-minded participants led to a reduction of stereotypes about individuals who hold different political views and improved perceived social cohesion in Germany. However, it did not lead to a change in political attitudes. In contrast, a conversation between like-minded participants neither reduced stereotypes nor had any impact on social cohesion. But, political attitudes changed by moving towards the boundary of our scale and away from the average opinion of our sample, which could be interpreted as a fortification of the political views. The results reveal a clear pattern: under the right circumstances, meeting people with substantially different political views can help to overcome negative consequences of rising polarization.

Chapter 2: "Stereotypes about Refugees - how motives mold peoples' stereotypes" examines the relationship between motives and belief formation. Studies show how the desire for a positive self-view leads individuals to adapt their beliefs about others' fairness and altruism. In this chapter, I investigate whether individuals go even further in their judgment of others to protect desired beliefs. To that end, I address the question of whether individuals, in situations in which they can benefit at the expense of refugees, adopt extremely negative beliefs or stereotypes about refugees to justify their selfish behavior? To answer this question, I conduct an online survey experiment using a representative sample of 902 German adults. Contradicting my hypothesis, I find that on average participants did not distort their beliefs about the refugees' image of women (*misogyny belief*). Using quantile regressions to detect variations in the treatment effect, I find that only those participants with extremely negative prior beliefs seemed to be pushed to state even more extreme beliefs in our treatment group. Further analysis revealed that a central design assumption of the experiment was violated, which while helping to explain the absence of average treatment effects, also provides suggestive evidence for another kind of belief distortion: ex-post rationalization. Participants who made a selfish decision before stating their beliefs about refugees, stated on average higher misogyny beliefs, perhaps as a way to justify the past decision that might otherwise reflect poorly on their character.

In contrast to the previous chapter, the last chapter of this thesis is concerned with the question of how individuals uphold a positive self-view even when

confronted with conflicting information. In joint work with Sven Heuser, **Chapter 3:** "**Self-serving attributions**" seeks to establish how individuals attribute feedback about themselves to external factors to maintain overconfident beliefs about themselves. Most feedback individuals receive is shrouded in multi-dimensional uncertainty, i.e. the feedback results from multiple underlying factors. To learn from such feedback, individuals must make attributions that connect the feedback to the various potential causes. Schneider, Hastorf, and Ellsworth (1979) famously wrote: "(...) We attribute success to our own dispositions and failure to external forces." In line with this claim, social psychologists frequently point towards self-serving motives as a way that individuals attribute feedback in such situations. We employ a multi-day lab experiment to present causal evidence on this phenomenon, studying the circumstances that facilitate self-serving attributions and assessing its consequences. In our experiment subjects received (noisy) feedback about their performance on an IQ-Test. The feedback did not solely depend on the performance in the IQ-Test, but also on an unknown state of the world. Unfortunately the data collection process was interrupted by the COVID-19 pandemic, rendering a large part of the analyses inconclusive, at least for now. Using the data we have at hand, we see no sign of motivated attributions towards the state of the world. In subsequent analyses, we showcase that updating under multi-dimensional uncertainty is a complex process that needs further studying. In particular, we show that the subjects' updating behavior differs depending on the sign of the feedback (negative vs. positive feedback), the randomly assigned type of the subject (privileged vs. discriminated against), and on initial overconfidence (biased prior beliefs).

A common thread of this thesis are the sources and consequences of distorted subjective beliefs and the question of how to thwart them. In the first chapter, I show that bringing together two contrary-minded individuals for a conversation can help to counteract negative consequences of rising political segregation by changing participants' beliefs about others and the state of the society. While biased beliefs were the point of origin in the first chapter, the following two chapters study the emergence of such beliefs. By exploring if individuals are willing to adopt extremely negative beliefs about an often marginalized group to justify selfish behavior, I test how far individuals are willing to go to protect desired beliefs. In the final chapter of this thesis, I suggest that individuals, in light of potentially damaging information, make self-serving attributions towards an external factor to uphold or even boost their self-view. It is my deep belief that studying how motives and societal tendencies, like political polarization, affect individuals' formation of beliefs can enhance our understanding of human behavior and is, therefore, a promising path for further studies.

References

Bishop, Bill. 2009. *The big sort: Why the clustering of like-minded America is tearing us apart*. Houghton Mifflin Harcourt. [1]

Schneider, David J, Albert H Hastorf, and Phoebe Ellsworth. 1979. *Person perception*. Random House. [3]

Sunstein, Cass R. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press. [1]

Chapter 1

Now we're talking: The impact of a political face-to-face conversation

Joint with Sven Heuser

1.1 Introduction

While academics still vividly debate the extent and form of political polarization in the masses, there are several well-established and disconcerting trends. We increasingly judge individuals that hold different political views as selfish, unintelligent, and even malicious (e.g., Iyengar and Westwood (2015), Boxell, Gentzkow, and Shapiro (2019), and Iyengar, Lelkes, Levendusky, Malhotra, and Westwood (2019)). At the same time, fewer individuals simultaneously hold both liberal and conservative views than in the past, dividing nations into distinct political camps (e.g., Dimock, Doherty, Kiley, and Oates (2014) and Gentzkow (2016)). Scholars like Sunstein (2018) and Bishop (2009) argue that one reason for these tendencies is the sorting of individuals into chambers in which they are not only isolated from differing views but also from those who hold these views. Following this argument, the observed lack of direct interaction with people outside the own chamber (see, e.g. Gentzkow and Shapiro (2011)) slowly leads to a society with irreconcilable differences in political attitudes and biased perceptions of people from other groups. Such a development poses a threat to Western democracies as the resulting inaccurate and polarized beliefs make it harder to compromise, while simultaneously diminishing trust and cooperation between groups in society (Downs, 1957; Becker, 1958).
If a lack of interaction is indeed the root of increasing societal devision, it remains to be seen whether increasing interactions across groups coud be the solution. This paper explores whether a face-to-face conversation can help to overcome the negative consequences of rising political polarization in Western societies. We investigate if a conversation has the potential to improve social coexistence by changing individuals' stereotypes about people with different political views and increasing perceived social cohesion. Further, we study if a political face-to-face conversation can bring in-

dividuals to question and adjust their political views. To answer these research questions, we look at individuals who registered for *Deutschland spricht*[1], a large scale program with the objective of bringing two contrary-minded individuals together for a private face-to-face conversation about politics. The program was launched by ZEIT ONLINE, the online appearance of one of Germany's largest newspapers (DIE ZEIT), and participants were recruited via a wide group of news outlets. During registration, individuals had to answer seven binary questions on different political topics.[2] Based on their answers to these questions, participants were matched with one another. If a matched pair decided to meet, they had a face-to-face conversation in a natural environment without any rules or external mediation. Complementing the program, we sent out surveys to all 19,138 registered participants before and after the conversations took place. Even though *Deutschland spricht* was not designed as a randomized controlled experiment, we can circumvent potential endogeneity issues and identify the intent-to-treat effects of a conversation by exploiting the pairing procedure of the program and carefully choosing treatment and control groups. We find that a face-to-face conversation can have effects on stereotypical beliefs about people with different political views, perceived social cohesion, and political attitudes. However, the size and the direction of the effects heavily depend on the constellation of the pair, i.e. whether a participant met a person with substantially different political attitudes or a rather like-minded person. Thus, throughout the paper we report the effects for the whole sample, the individuals who were matched to a *like-minded* partner and the individuals who were matched to a *contrary-minded* partner.[3]

Our first set of results pertains to stereotypes about people who hold different political beliefs. The design of *Deutschland spricht* allows us to individually define such a person by referring to people who answered the seven political registration questions very differently from oneself. As outlined above, one disconcerting trend in many western democracies is that individuals view others that do not share the same convictions as increasingly ill-informed, unworldly, and morally questionable. To see if a face-to-face conversation can overcome such tendencies, we elicited four stereotypical beliefs that aim to capture perceptions concerning the incompetence and otherness of people who hold contrasting political beliefs. Further, we gathered data on the willingness to have a person with substantially different political beliefs in one's social environment. Following the seminal work by Allport, Clark, and Pettigrew (1954), research on intergroup contact suggests that interactions between peo-

1. Translation: Germany talks
2. The questions were chosen by ZEIT ONLINE with the aim to be as controversial as possible.
3. We define a *contrary-minded* partner as a person that answered more than three of the seven binary political registration questions differently. Thus, if a participant meets a *contrary-minded* person there exists more disagreement than agreement between them. A *like-minded* partner is a person that answered three or less questions differently.

ple from different groups should help to overcome prejudice and stereotypes.[4] Thus, a face-to-face conversation with a person from a different political camp should help to overcome the increasingly negatives views individuals hold about people with different political opinions. It should be emphasized that an individual must meet a sufficiently contrary-minded person in order for real change to occur. Conversations with partners who have similar political attitudes should not affect the beliefs about those with opposing views. In line with our hypothesis, we find that stereotypes declined if participants met *contrary-minded* partners. Collapsing the four beliefs into two stereotypes *Incompetence* and *Otherness*, we find a significant reduction of 0.329 (at a 1% level) and 0.201 standard deviations (at a 5% level), respectively. This goes along with a (insignificantly) higher willingness to have people with opposing political views in one's social environment. In the case of a *like-minded* partner, we do not find any effect. If anything, there is a tendency towards reinforcement of stereotypes and reduction of willingness.

The second set of results focuses on social cohesion, proxied by two variables capturing how fellow German citizens are perceived. The first variable measures whether German citizens are considered trustworthy (*Trust*). The second one is the belief of how much German citizens generally care about the wellbeing of others (*Care*). Following the rationale from above, we expect improved social cohesion through contact. Contact with a partner that does not share the same beliefs should be more revealing about the society as a whole as it presents insights into a part of society the participants do not frequently interact with and might have biased beliefs about. Thus, we should see greater effects for contrary-minded pairs. We find that the conversation significantly raised both social cohesion proxies by 0.185 (*Trust*) and 0.168 of a standard deviation (*Care*) in the whole sample (both at a 1% level). As expected, the effects are driven by people who met *contrary-minded* partners (0.221 and 0.283 standard deviations, respectively) while meetings with like-minded participants show a similar, albeit insignificant tendency.

Our third set of results looks at the interventions impact on political attitudes and examines whether a face-to-face conversation has the potential to sway the participants' attitudes and bring individuals in their views closer together. Exchange of new and confronting, but also familiar and similar viewpoints between the two partners during the discussion may lead to an adjustment of their political attitudes. Research on deliberation suggests that meetings with like-minded partners should lead to a mutual reconfirmation of views.[5] In contrast, it is unclear what to expect when contrary-minded people talk. Some research suggests that a confrontation

4. See, for example Boisjoly, Duncan, Kremer, Levy, and Eccles (2006), Burns, Corno, and La Ferrara (2015), Broockman and Kalla (2016), Finseraas, Hanson, Johnsen, Kotsadam, and Torsvik (2016), Paluck (2016), and Rao (2019). Pettigrew and Tropp (2006) provides a meta-analysis.

5. When individuals deliberate with like-minded people, beliefs and attitudes are reconfirmed, i.e. conservatives become more conservative and liberals more liberal. See, for example, Sunstein (1999), Schkade, Sunstein, and Hastie (2007), Glaeser and Sunstein (2009), and Sunstein (2009).

with differing viewpoints leads to convergence of attitudes (Gutmann and Thompson, 1998) while other studies do not (Wojcieszak, 2011). For this investigation, we elicited the level of agreement to eleven political attitude statements on a seven point Likert scale before and after the meeting. This enables us to collapse the eleven single attitudes into one overall attitude and analyze its change. The first step of our analysis is to investigate whether the discussion caused a closer examination and an adjustment of the own views. We find that the conversation about politics led to a significant change in political attitudes of 0.179 standard deviations (at a 1% level). This effect is driven by conversations between *like-minded* persons (0.212 standard deviations). For *contrary-minded* partners we find a positive but insignificant effect (0.166 standard deviations, significant on a 10% level). In a next step, we explore the direction of the adjustment by addressing both whether individuals move towards the boundary of our scale and whether they move towards the average pre-conversation opinions in the impact sample. We argue that a movement towards the boundaries and away from the average opinion can be interpreted as a movement to more extreme views. We observe that the direction of adjustments differ depending on the pair constellation. As hypothesized, a conversation with a *like-minded* partner moved participants 0.261 standard deviations closer to the boundaries of our scale (significant at a 1% level). Moreover, we observe a significant divergence away from the average opinion of our sample (0.213 standard deviations, at a 5% level). In contrast, meetings with a *contrary-minded* partner did not move the overall attitude in one consistent direction. However, we observe a significant convergence towards the average opinion (-0.193 standard deviations, at a 5% level).

The overall picture of the effects provides important insights under which circumstances face-to-face discussions are desirable. Meeting someone from outside the own group clearly has a positive impact on stereotypes and perceived social cohesion. Moreover, we do not see any sign for further divergence in political attitudes, but if anything, a tendency to move closer together. On the other hand, meetings with persons from the own group merely entrench political attitudes and drive people further away from those with differing views. Thus, from a policy perspective discussions between different political groups should be facilitated to fight the segregation of political camps. This seems even more urgent in the light of the fact that staying inside the own echo chamber widens the gap between the groups.

We contribute to the existing research in three ways. First, as pointed out by Paluck (2016), there is a shortcoming of studies that us adults to test the causal effect of real-world interventions in the intergroup contact literature. To the best of our knowledge, this is the first large field study that uses adults to present causal evi-

---

Further support for this idea comes from the literature on bounded rational belief formation. As individuals neglect the underlying correlation between signals, they end up with biased beliefs (see, e.g. DeMarzo, Vayanos, and Zwiebel (2003), Eyster and Rabin (2014), and Enke and Zimmermann (2017))

dence on the impact of a face-to-face conversation with a person who has contrary political views. Second, we contribute to research on deliberation. The studied conversations are private and are neither mediated nor are the topics determined by experimenters. Thus, by examining discussions in a unique and natural field setting, this paper contributes to the literature on the effects of political discussions and their importance for the democratic process. Third, we also build on research on polarization. As previously discussed, there is a vivid debate between academics about whether we are living in polarized times or not. While some argue that we observe mass polarization (see, e.g. Abramowitz and Saunders (2008)), others argue that it is a myth (see, e.g.Glaeser and Ward (2006) and Fiorina and Abrams (2008)).[6] However, there is a great deal of evidence that we are living in a time characterized by strong partisanship (Gentzkow and Shapiro, 2010; Flaxman, Goel, and Rao, 2016; Chen and Rohla, 2018; Gentzkow, Shapiro, and Taddy, 2019), affective polarization[7] (see, e.g. Iyengar, Sood, and Lelkes (2012), Iyengar and Westwood (2015), Boxell, Gentzkow, and Shapiro (2019), and Iyengar, Lelkes, et al. (2019)), elite polarization (Druckman, Peterson, and Slothuus, 2013) and increasing consistency in political attitudes (Center, 2014; Gentzkow, 2016). We contribute to this literature by investigating under which circumstances a face-to-face conversation may help to mediate the above described tendencies.
The paper proceeds as follows. In Section 1.2, we give an overview about the program *Deutschland spricht* and our measures. Section 1.3 presents our empirical strategy and descriptive statistics. Section 1.4 reports our results and discusses robustness. Section 1.5 concludes.

1.2 The Intervention

In this section, we give an overview of the program *Deutschland spricht* and introduce our measures.

1.2.1 Overview

Deutschland spricht is a program organized by ZEIT ONLINE. It was launched in 2017 with the general objective of bringing pairs of individuals that hold contrary political opinions together for a face-to-face conversation about politics. In 2018 it was conducted in cooperation with several other large German news outlets.[8] Com-

6. As noted by Gentzkow (2016) much of the disagreement stems from different definitions of polarization. While most of the academics focus on attitude polarization in the US, some researchers focus on cultural polarization (Desmet, Ortuño-Ortın, and Wacziarg, 2017; Bertrand and Kamenica, 2018), moral values (Enke, 2018) or look at countries outside the US (Munzert and Bauer, 2013).

7. Affective polarization is defined as the mutual dislike between partisan groups (Lelkes, 2016).

8. The 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

plementing the program, we sent out questionnaires to all participants before and after the meetings took place. Figure 1.A.1 summarizes the timeline.

The news outlet recruited participants using their respective media platforms. In order to register for the program, individuals had to answer seven binary questions about contemporary political topics. Among others, they had to state whether German inner cities should be car-free and whether Muslims and non-Muslims cohabit well in Germany. Table 1.A.1 lists all seven questions, henceforth referred to as *political registration questions*. In addition to these questions, applicants had to state further basic information about themselves and answer five non-political free response questions.[9]

After the registration phase closed, a matching algorithm attempted to pair each one of the registered and validated 19,134 participants.[10] The goal of the algorithm was to match each participant with another individual who gave as different answers to the *political registration questions* as possible and lived within a 20 km radius.[11] The algorithm ultimately matched 91% of participants successfully. The remaining 9% of the participants were informed that no partner had been found. Each successfully paired individual received an email introducing the matched partner. This email contained the partner's first name, age, gender, the answers to the five free response questions and - moreover, a list of the *political registration questions* the partner had answered differently. Based on this information, the participants could decide whether they wanted to accept the suggested partner or not. After the first person had accepted the partner, another email was sent out to notify the second participant that her assigned partner is willing to meet. When both partners had confirmed the match, contact was established by giving out the respective email addresses.[12] From this point on, the organizers of *Deutschland spricht* played no further role in organizing the meetings. Participants had to organize time and location of the meeting themselves. However, the suggested and officially communicated date of the conversations was September 23, 2018 and most pairs met on that date. The meetings were not observed, i.e. there was no third party moderating or observing the discussion and no rules or topics of discussion were predefined. Thus, the meet-

9. The additionally requested information included name, gender, age and zip code. The five free response questions asked participants about themselves, their hobbies and dislikes (see Table 1.A.2). Unfortunately, neither the name nor the answers to the five free response questions were made available to us.

10. Initially, *Deutschland spricht* received around 28,000 applications. However, approximately 8,000 applications turned out to be invalid and were removed by the organizers of *Deutschland spricht*.

11. The main objective of the algorithm was to match as many participants as possible, while fulfilling the following two conditions: (1) The matched partner had to be located in a 20 kilometer perimeter (the locations were the geographical centers of the respective zip codes). (2) Given the fulfillment of (1), the differences in the answers (number of differently answered political registration questions) was maximized.

12. Among the matched pairs, 45% were pairs where both partners accepted, 39% were pairs where only one partner accepted and 16% were pairs where none of the partners accepted.

ings were natural in the sense that they were private face-to-face discussions among two partners who could discuss whatever they pleased.

As a complement to the program, we sent out two emails to all registered participants. The emails contained personalized links to our pre- and post-survey and were sent out five days prior and eight days after the suggested day of the conversations, respectively.[13] To maximize survey completion rates, we kept both surveys relatively short.[14] As a consequence, most sociodemographic control variables were only elicited in the pre-survey. In the subsequent subsection, we describe in detail which outcome and control measures were elicited at which point in time.

In sum, we have data from three different sources: For all 19,138 registered participants of *Deutschland spricht* we have the data obtained during the registration process including time stamps and identifiers of the matched partners. Additionally, 5,677 participants filled out the pre-survey while 4,200 respondents completed the post-survey. We have all data points for 2,645 participants.

1.2.2 Measures

Outcome Variables

To answer whether a face-to-face conversation has the power to alter participants' political attitudes, stereotypical beliefs about people with different political views and perceived social cohesion in Germany, we elicited a variety of dependent variables in both surveys. Table 1.1 summarizes our outcome variables.

Table 1.1. Outcome Variables

Variable	Statement	Answer Scale
Political Attitudes		
Border Control	Germany should implement stronger border controls.	*0 = Strongly disagree* to *6 = Strongly agree*
Metoo	The public debate about sexual harassment and metoo caused something positive.	*0 = Strongly disagree* to *6 = Strongly agree*
Meat Tax	Meat should be taxed more to reduce its consumption.	*0 = Strongly disagree* to *6 = Strongly agree*
Car-free Inner-Cities	German inner-cities should be car-free.	*0 = Strongly disagree* to *6 = Strongly agree*
Muslims	Muslims and Non-Muslims cohabit well in Germany.	*0 = Strongly disagree* to *6 = Strongly agree*
German Development	Germans are worse off than 10 years ago.	*0 = Strongly disagree* to *6 = Strongly agree*
Trump	Donald Trump is good for the USA.	*0 = Strongly disagree* to *6 = Strongly agree*
Same-Sex marriage	Marriage should only be possible between a men and a women.	*0 = Strongly disagree* to *6 = Strongly agree*
EU-States	Germany should deepen its cooperation with other EU countries.	*0 = Strongly disagree* to *6 = Strongly agree*
Media	Altogether, German media are trustworthy.	*0 = Strongly disagree* to *6 = Strongly agree*
Income Tax	To reduce the gap between rich and poor, the maximum tax rate for top earners should be increased.	*0 = Strongly disagree* to *6 = Strongly agree*
Stereotypical Beliefs		
Otherness (Part 1)	This person has completely different moral values.	*0 = Doesn't apply at all* to *6 = Fully applies* (rev.)
Otherness (Part 2)	This person leads a completely different life.	*0 = Doesn't apply at all* to *6 = Fully applies* (rev.)
Incompetence (Part 1)	This person is badly informed.	*0 = Doesn't apply at all* to *6 = Fully applies* (rev.)
Incompetence (Part 2)	This person is incapable of understanding complex contexts.	*0 = Doesn't apply at all* to *6 = Fully applies* (rev.)
Willingness to Interact	I don't want this person to be in my personal environment.	*0 = Doesn't apply at all* to *6 = Fully applies* (rev.)
Social Cohesion		
Trust	You can trust most people in Germany.	*0 = Strongly disagree* to *6 = Strongly agree*
Care	Most people in Germany do not care about the wellbeing of their fellow citizens.	*0 = Strongly disagree* to *6 = Strongly agree*

The table lists all outcome variables including a translation of the original formulations. The *Stereotypical Beliefs* were elicited directly after participants answered the first seven political attitude questions. We asked participants to picture a person that gave *very different* answers to the seven political attitude questions and to state their beliefs about this person. The last column shows the corresponding scales. We reversed the scale for stereotypical beliefs during analysis for interpretational purpose.

13. The pre-survey was sent out more than a week after the participants had received the first email introducing their suggested partner. Hence, the majority of the individuals who wanted to meet had already confirmed their partner and email addresses had already been exchanged by the time participants filled out the pre-survey.

14. The average durations were 14 minutes for the pre- and 12.5 minutes for the post-Survey.

Political Attitudes. Participants were asked to state the extent to which they agree with different political statements on a seven point Likert scale. See Table 1.1 for an overview. The first seven of the eleven questions were, apart from the transformation from questions into statements, identical to the *political registration questions*. However, participants could now indicate their (dis-)agreement on a finer level. In addition to the seven questions, we elicited four other, more general political attitudes. Based on these attitudes, we create a variety of variables for our analysis. The underlying idea is to take all eleven attitudes together and interpret the eleven dimensional vector as the *overall* political attitude.

Absolute change - The measure *PA_change* helps us to explore if there was any movement in participants' overall political attitude. We define PA_change_i as the Euclidean distance between individual i's overall attitude in period 2 (after the conversation) and period 1 (before the conversation):

$$PA\_change_i = \sqrt{\sum_{a=1}^{11} (Y_{ai2} - Y_{ai1})^2}$$

where Y_{ait} denotes individual i's level of agreement to attitude a in the Post-Survey (t=2) and the Pre-Survey (t=1). The eleven attitudes are the political attitudes from Table 1.1. It should be noted that PA_change_i is neutral towards the direction of movement, i.e. it merely captures the magnitude of change.

Change towards the center of our scale - The measure $PA\_change\_center_i$ indicates whether someone moved towards or away from the center of our scale (a vector of 3s), we define:

$$PA\_change\_center_i = \sqrt{\sum_{a=1}^{11} (Y_{ai2} - 3)^2} - \sqrt{\sum_{a=1}^{11} (Y_{ai1} - 3)^2}$$

where Y_{ait} denotes individual i's level of agreement to attitude a in the Post-Survey (t=2) and the Pre-Survey (t=1). The first term is the Euclidean distance of i's attitude to the center point (vector of 3s) in the Post-Survey (t=2), while the second term is the respective Euclidean distance in the Pre-Survey (t=1). Thus, $PA\_change\_center_i$ reflects the change in the distance to the center 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 the average opinion - To see whether the meeting moves people closer to the average opinion of the population, we construct the variable

PA_change_average$_i$. The variable measures whether an individual converged to the average pre-meeting overall attitude of our impact sample:

$$PA\_change\_average_i = \sqrt{\sum_{a=1}^{11} (Y_{ai2} - \overline{Y}_{a1})^2} - \sqrt{\sum_{a=1}^{11} (Y_{ai1} - \overline{Y}_{a1})^2}$$

where Y_{ait} denotes individual i's level of agreement to attitude a in the Post-Survey (t=2) and the Pre-Survey (t=1). $\overline{Y}_{a1}$ is the average level of agreement to attitude a of all participants in the impact sample in the Pre-Survey. The two terms reflect the distance to the average pre-meeting opinion after and before the meeting took place. In sum, *PA_change_average_i* denotes whether someone moved towards (*PA_change_average_i* < 0) or away from (*PA_change_average_i* > 0) the average pre-meeting opinion or none of the two.

Stereotypical Beliefs. In order to study how the conversations affected stereotypes about people from different political camps, we collected data on participants' beliefs about a person whose answers to the political attitude statements differed considerably from those of the participant. The five questions are reported in Table 1.1.

Otherness - In two separate questions, individuals were asked to state their beliefs about the degree to which the other person's way of life and their moral values differ from their own. As both questions capture a very similar concept (the difference between oneself and the other person), we combine them by running a Principal Component Analysis. We call the derived principal component *Otherness*.

Incompetence - Analogously to *Otherness*, we derive the variable *Incompetence* by performing a PCA with the beliefs about the other person's cognitive abilities and how poorly informed the other person is.

Willingness to interact - We elicited participants' level of agreement to the statement that they do not want to have a person that gave *very different* answers to the political registration questions in their social environment. Reversing the scale yields a participant's *willingness to interact* with a person who holds opposing beliefs.

Social Cohesion. Participants answered two questions about perceived social cohesion in Germany. The variables measure two different components of how participants perceive their fellow German citizens: The first one elicits the belief about how trustworthy German citizens are in general (*Trust*). The second one asks whether German citizens generally care about the wellbeing of others (*Care*).

In addition to the variables above, we elicited several outcome measures which we do not focus on in the paper. First, our surveys contained measures for the relative size of the Muslim population in Germany and the total number of refugees in

Germany as of 2017. We asked these questions to see whether people are objectively better informed after the meeting. However, we know that participants only rarely talked about these numbers[15] and therefore do not expect that the conversation had an impact in this dimension. Further, we elicited perceived social acceptance. The participants in our impact sample have in general a very high perceived acceptance leaving very little room for improvement due to the meeting. Last, we asked people about their beliefs about attitudes of the average AfD and Green voter, respectively. Unfortunately, we do not have sufficient data to analyze the effects of the meeting on these beliefs.[16]

Non-Outcome Variables

Political distance, contrary-minded & like-minded. The variable *political distance* denotes the number of the binary political registration questions a participant answered differently from her partner. *Political distance* takes values between 0 (all political registration questions were answered identically) and 7 (all political registration questions were answered differently). *Political distance* is a proxy for the differences in political attitudes between the partners. For example, a pair with *political distance* of one answered six of the seven political registration questions identically. Hence, they seem to be rather like-minded in their political attitudes. Throughout our analysis we will call pairs with a *political distance* between 0 and 3 *like-minded* and pairs with a *political distance* between 4 and 7 *contrary-minded*.

First movers, second movers. Using the timestamps of the emails from both partners, we construct if and when each participant accepted a match and whether she was the first person in the pair. We call the partner who accepted the match first *first mover* and, analogously, the partner who accepted second *second mover*.

Control Variables. Additional to the variables elicited during registration, we gathered more control variables about the participants in our pre- and post-survey. In the pre-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 position on a political left-right spectrum, and the party they would vote for. In the post-survey we elicited income and marital status.

15. Only 8.5% of the participants talked about the number of refugees and only 9 % about the percentage of Muslims living in Germany.

16. We have relatively few observations that contain the party preferences of both partners, making it impossible to identify those cases in which there was intergroup contact.

1.3 Empirical Strategy

In this section, we discuss the empirical strategy used to identify the causal impact of a political face-to-face conversation and present descriptive statistics.

1.3.1 Identification

Potential Challenges: Selection Biases. As we do not have a classic experimental setup, it is imperative to pick our treatment and control group carefully and take care of various (potential) selection biases to avoid endogeneity. To start with, one fundamental concern is that individuals who select themselves into the conversation are systematically different from individuals who do not. For example, those who select themselves into the conversation may be inherently more receptive to differing viewpoints. To limit potential problems arising from these concerns, we restrict the sample to only those participants who accepted their partners first (*first movers*), thereby allowing us to compare only those participants with a high willingness to participate in a political face-to-face conversation. Within this sample, we compare participants who were also accepted by their partner (treatment group) with those who were not accepted (control group). Because the decision of the *second mover* essentially determines who will and will not have a conversation, i.e. who will be in the treatment and who will be in the control group, two further conceivable sources of selection biases arise: (i) A selection bias caused by *second movers'* acceptance being contingent on characteristics of *first movers* and (ii) a selection bias stemming from the assignment of partners.
(i) refers to the (valid) concern that *second movers* make their choice depending on the characteristics of the *first movers*. In this case, specific types of *first movers* are more likely to be in the treatment group than others. As soon as these types of *first movers* also show different behavior regarding the outcome variables, we face endogeneity.
(ii) refers to the concern that specific types of *first movers* are matched with higher probability to individuals who idiosyncratically accept more often. Consider the following example: Suppose there are two types of *first movers*, A and B. Type A has strong prejudices and lives in a city where individuals are generally less open towards other individuals. Type B, in contrast, has few prejudices and lives in a city where individuals are highly willing to get to know others. Because of the geographical restraint on matching, B types are more likely to meet other accepting types and vice versa for A types. As a consequence, B types are more likely to be in the treatment group than A types, leading to a biased estimation of the impact of a political face-to-face discussion on prejudices. One would see a reduction of prejudices and wrongly attribute it to the treatment, even though in reality it stems from the overrepresentation of Type B in the treatment group.
We take the following steps to solve the above problems: First, we control for all the

available information the *second mover* knows about the *first mover* when deciding whether to accept or not. This way, we aim to make the choice conditionally independent of the *first movers'* characteristics (selection bias (i)). Second, we exploit the matching algorithm to tackle selection bias (ii). Participants were matched solely based on their answers to the political registration questions and the constraint that the partners' places of residence were within 20 kilometers of each other (measured by the centers of the zip codes). Adding regional fixed effects and the answers to the binary political registration questions allow us to account for the endogeneity stemming from the selection bias (ii). The estimation strategy is discussed in more detail below.

Impact Sample and Data Restriction. We now briefly outline which data we use for our analysis. As discussed in the paragraph above, we only consider *first movers* in order to avoid selection bias. Further, we restrict our sample to those *first movers* who answered both surveys, as most of our required controls were elicited in the pre-survey, while the conversations' impact can only be detected in the post-survey. Lastly, for most of our analysis we will only use outcome measures from the post-survey. The reason for this is that the pre-survey was unfortunately sent out too late, as 97% of the participants in the treatment group were already accepted by their partner (and thus had already learned their treatment assignment) when filling out the survey.[17] By then, most of them probably had already established contact with their partners to set up a date and location to meet. This is particularly problematic when investigating the effect on stereotypes, trust or similar outcomes, as these may easily be influenced by any (first) contact. Besides, learning the treatment status itself might have affected the answers. As a consequence, we drop the outcome measures from the pre-survey in almost all our analyses. The only exception to this is when we look at whether political attitudes changed due to the discussion, as this investigation is only possible when including both points in time. Consequently, these analyses are only valid under the assumption that political attitudes are not affected by either learning the treatment assignment or first email contact to arrange the meeting and should thus be looked at with caution.

Estimation. We estimate the following model using OLS:

$$Y_i = \alpha + \beta * Treat_i + \gamma * Controls_i + \epsilon_i \tag{1.1}$$

where Y_i denotes our outcome variable from the Post-Survey. The dummy $Treat_i$ indicates whether first mover i was accepted by her partner (the second mover) or not and ϵ_i is an individual specific error term. As discussed in the previous paragraph, $Controls_i$ contain control dummies to prevent selection issues. β measures

17. The surveys were sent out by the organizers of *Deutschland spricht*. We were merely responsible for the content.

the intent-to-treat effect of a political face-to-face discussion. We have very high compliance rates (100% for the control group and 87% for the treatment group) suggesting that the the average treatment effect of the discussion is very similar to our intent-to-treat measure.[18]

For each outcome measure we always report the results of three different specifications of (1.1) varying the set of control dummies. In every specification, we include region fixed effects and the answers to the political registration questions to alleviate concerns arising from potential selection bias (ii).[19] To achieve conditional independence of the *second mover's* acceptance decision (probability of being treated) from individual i's characteristics (potential selection bias (i)), we additionally control for as much information the second mover knows about the first mover as possible. As described in the previous section, participants know their partner's first name, gender, age, as well as their answers to both the political registration questions and the five open questions when deciding whether to accept. We have available to us the gender, age and the political registration questions for all *first movers*, but we do not have the answers to the five open questions nor the surname. Thus, the first, most basic specification contains only the "hard facts", i.e. the information we have and which we can easily control for in our regression.[20] We call this set of controls *Basic Controls*.

In the next step, we extend the set of controls to include the remaining information the *second movers* know about the *first movers*. In particular, we add variables that might be visible either through the surname or the answers to the open questions and which might be correlated to our outcome variables. First, we include dummies for income and education, to control for socioeconomic status.[21] Information about socioeconomic status might especially be visible in the surname and the answers to the open question regarding the participant's job. Second, we control for migration background. Whether someone has a migration background is potentially visible through the surname and might be correlated to prejudices or political attitudes. Last, we add dummies for political self-classification (from left to right) and voting behavior (party) because these factors are correlated to most of our outcome variables and also likely contained in the answers to the open questions. We call this additional set of further controls *Name & Political Controls*.

18. We will discuss this in more detail in Section 1.4.4,

19. For region we use dummies for two-digit zip codes, a regional aggregation of higher order zip codes (i.e. zip codes with more digits) instead of five-digit zip codes which were used by the matching algorithm because finer scales increase the number of control variables exorbitantly: We have 95 different two-digit zip code values while three-digit zip code would yield 541, four-digit zip codes 1049 and five-digit zip codes 1531 dummies.

20. We divide age into intervals: 18-25, 26-35, 36-45, 46-55, 56-65, 65+.

21. We only elicited monthly income categories. We create dummies for each of these categories: Prefer not to say, 0 - 800 Euros, 800 - 1500 Euros, 1500 - 2200 Euros, 2200 - 3300 Euros, More than 3300 Euros, I don't know.

Our third specification tries to capture even more of the information contained in the answers to the open questions. The *Open Question Controls* contain dummies for religion, piety, marital status and social environment. While it is possible that these variables are visible through the open questions, we do not necessarily believe them to be correlated with our outcome variables.
Hence, under the assumption that our controls lead to conditional independence of the *second mover's* acceptance decision from the *first mover's* characteristics and, additionally, sufficiently erase selection bias (ii), β measures the intent-to-treat effect of a political face-to-face discussion. In the body of the paper, we focus on the specification with all controls, henceforth referred to as our *main specification*.

Heterogeneous Treatment Effects. We argue that whether a participants meets a like- or a contrary-minded partner is an important distinction that would likely lead to different effects. Using *political distance*, a variable we have for every participant, as a proxy for dissimilarity in political opinions within a pair, we divide the sample into two fairly balanced subsamples: The first subsample consists of participants that had a *political distance* of 4 or higher, i.e. the two partners answered more than half of the political registration questions differently from each other. The second subsample consists of the remaining participants, i.e. all participants who answered more than half of the political registration questions identically. Synonymously, we use the terms *like-minded* partner and *contrary-minded* partner throughout the paper. We run separate analyses for the whole, the *contrary-minded* and the *like-minded* sample to show the complete picture of the effects.

1.3.2 Descriptive Statistics

Table 1.B.1 quantifies the composition of our impact sample, which is composed of 1523 participants. Compared to the German population, our sample is relatively well-educated, male, politically left-leaning and without migration background, but similar with respect to age and place of residence.[22] This reflects the fact that the program and thus the recruitment was mainly done by predominantly left-liberal, but Germany-wide sold newspapers.
Within the impact sample, there are 969 subjects in the treatment and 554 in the control group. To see whether the likelihood to be treated is conditionally independent of the individuals' characteristics, we executed a variety of balance checks. We run our main specification using variables we collected but not focus on in this paper and variables that are not affected by the treatment assignment (e.g. all the political

22. The comparison is based on data of the German Federal Office in Statistics: Migration background (whole population: 25 %, impact sample: 10%), Percentage male (whole population: 48%, impact sample: 63%), Education - University Degree (whole population: 17.6%, impact sample: 67%), age (whole population: 44.4, impact sample: 47.43)

attitude measures from the pre-survey) as dependent variables.[23] Table 1.B.2 summarizes the results for the whole, the *contrary-minded* and the *like-minded* samples. In the whole sample only two out of 18 variables are significant, and in the two subsamples only one and zero variables, respectively, are significant, suggesting that in all our samples treatment and control groups are conditionally balanced.

Although participants in our impact sample are rather homogeneous in their party and political preferences[24], there still exists a fair amount of dissonance among them. Figures 1.B.1 and 1.B.2 reveal that there was considerable variation in the participants' positions in the pre-survey, leading to substantial heterogeneity in the political attitudes of the pairs. This is also reflected in the distribution of the *political distance* within the pairs in our sample (see Figure 1.B.3). Overall, the distribution is fairly balanced with respect to our split into like- and contrary minded pairs. 49% of the participants were matched with a *contrary-minded* person, while 51% were assigned to a *like-minded* person. On average, matched partners answered 3.5 out of the seven political registration questions differently.

As there were no guidelines regarding content and procedure of the conversation nor any control by the organizers of *Deutschland spricht*, we asked participants about their experiences and what they discussed during their meetings. As shown in Figure 1.B.4, the conversation centered around the topics of the seven political registration questions. On average, the conversation lasted 140 minutes and an overwhelming majority of the participants reported that it was a pleasant experience.[25]

1.4 Results

This section presents the effects of a political face-to-face discussion depending on whether a person met a like- or contrary-minded partner. In the first subsection, we show how the conversation affected participants' stereotypes about people that hold different political beliefs. Second, we outline the impact on perceived social cohesion, and last we provide effects on participants' political attitudes.

In the body of the paper, we present plots of the treatment coefficients of our main specification with all controls and their corresponding 95% confidence intervals

23. In the last paragraph of Section 2.2, we listed variables that we do not focus on in this paper. The variables are the two estimates, the social acceptance question and the beliefs about Green and AFD voters.

24. We observe that a majority of the individuals identify themselves as left of center on a political spectrum and the liberal and eco-friendly Green party would receive 47% of the votes. To put this in perspective, in a representative poll by Forsa from the day the registration for *Deutschland spricht* started, the Green Party reached 16%.

25. 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 (Participants had to state how much a statement applied to their conversation on a seven point Likert Scale (0 - 6). The reported percentages are for those who reported either (5) *agree* or (6) *strongly agree*).

using three different samples: (i) The whole sample, (ii) the *like-minded* sample, and (iii) the *contrary-minded* sample. (i) contains all subjects, (ii) only those who were matched to a *like-minded* partner, and (iii) only those who were matched to a *contrary-minded* partner. The outcome variables are normalized such that the control group has mean zero and standard deviation one. The corresponding regression tables, including the results of the other specifications, can be found in Appendix 1.C.

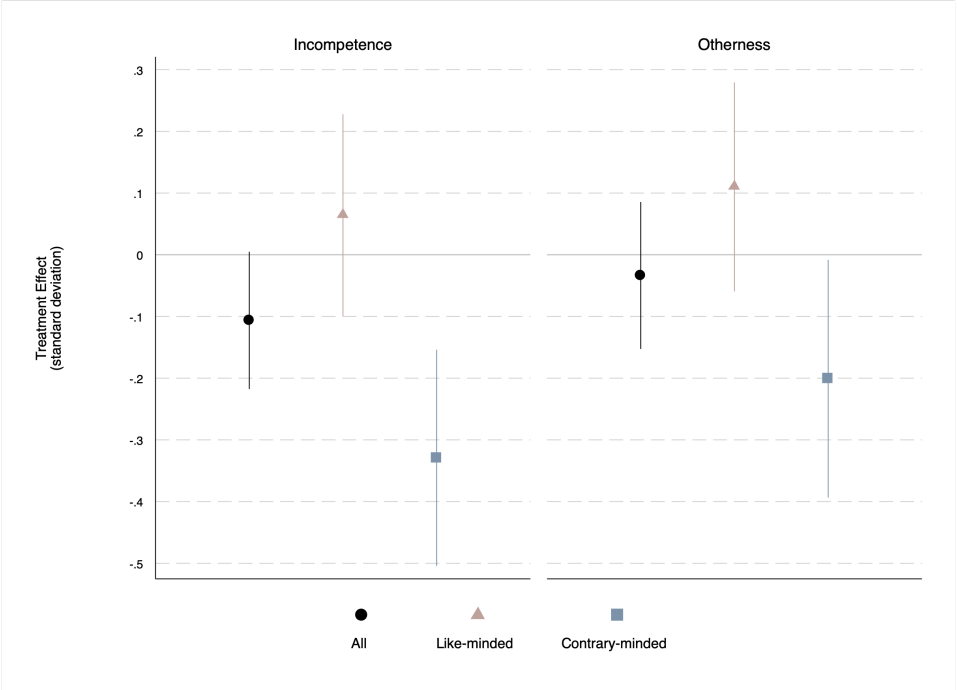
1.4.1 Effect on Stereotypes

One important question in the context of political face-to-face conversations is whether they can help people to see those with contrasting opinions through a different lens and therefore help to reduce rising affective polarization. As outlined in the introduction, research on intergroup contact suggests that we should expect a positive effect, but only if someone met a person with sufficiently distinct views. Thus, in the context of our setup we should see a reduction of stereotypes if a person was matched with a *contrary-minded* partner and no reduction if she was assigned to a *like-minded* partner.

Stereotypes: Otherness and Incompetence. Figure 1.1 shows the effects on stereotypes *Otherness* and *Incompetence* for the whole, the *like-minded* and the *contrary-minded* sample. The figure reveals a clear pattern: A face-to-face conversation with a *contrary-minded* person significantly lowered both of the participants' stereotypical beliefs. *Incompetence* was reduced by 0.329 and *Otherness* by 0.201 standard deviations. At the same time, a conversation with a person that holds similar viewpoints failed to reduce the stereotypical beliefs. The positive coefficient suggests that, if anything, stereotypical beliefs are reinforced. Taking both samples together, there is a (insignificant) tendency towards a reduction. The exact values of the corresponding regressions and the results of the further specifications are reported in Tables 1.C.1 and 1.C.2. Both tables show that our findings do not depend on the specification we use.[26]

Willingness to Interact. Figure 1.2 presents the effect of the conversation on the willingness to interact with people who have a very different political opinion. In line with the previous finding, we see that the outcome depends on the pair composition. Even though none of the coefficients are significant for any of the three samples, we see that there is a positive coefficient for *contrary-minded* pairs (higher willingness to interact) while it is negative for *like-minded* pairs (lower willingness to interact). Considered individually, the results are not indicative for meaningful change, but they fit in the pattern we see for the stereotypes: Meeting a *contrary-minded* partner

26. We find the same pattern when looking at the four stereotypes separately. See Appendix 1.E for the respective Tables.



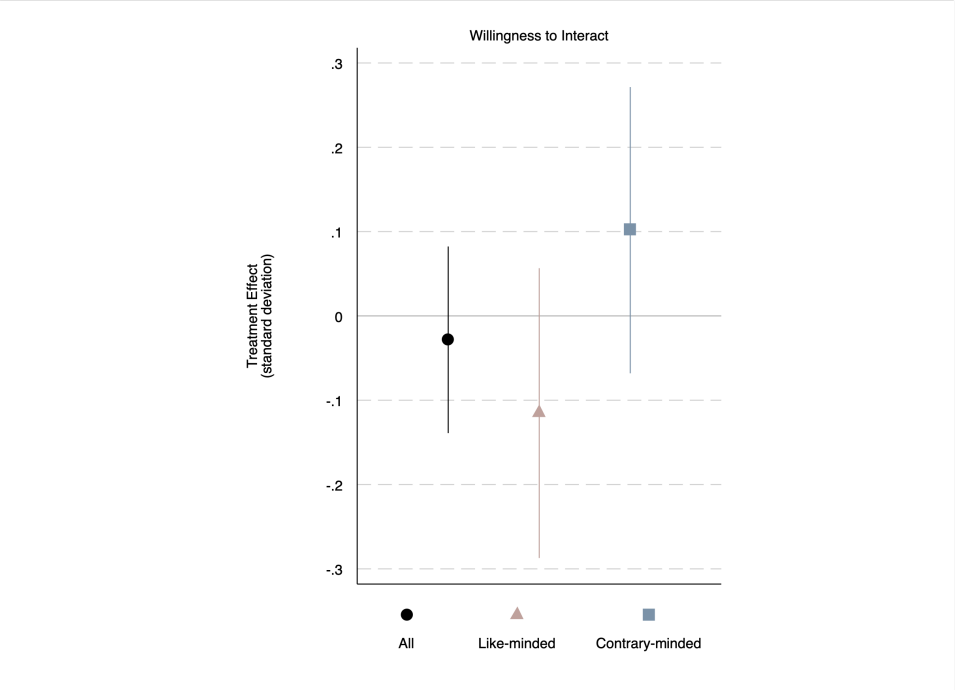
The figure shows treatment effects on (standardized) stereotypical beliefs *Otherness* and *Incompetence*. It depicts the estimated treatment coefficients of our main specification and the respective 95 % confidence interval. The point reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the like-minded subsample (*Political Distance* 0-3) while the square and the corresponding line report the results for the contrary-minded subsample (*Political Distance* 4-7).

Figure 1.1. Effect on Stereotypes

had a positive impact while having a discussion with a *like-minded* partner led to a negative impact. Table 1.C.3 shows the exact results of all specifications.

Result 1. *Face-to-face conversations with contrary-minded partners significantly reduced stereotypes about people with different political opinions. Conversations with like-minded partners did not have a significant effect.*

The results in this section confirm the hypothesized pattern and paint a coherent picture. As expected, a face-to-face conversation with a person that holds substantially different political attitudes can have positive effects that lead to lower stereotypes. This positive effect is confirmed by the tendency of a higher willingness to interact with outgroup members due to the meeting. In contrast, meeting a *like-minded* person does not have an effect. If anything, we find signs for a negative, exacerbating impact of the discussion.



The figure shows the treatment effect on the (standardized) willingness to interact. It depicts the estimated treatment coefficient of our main specification and the respective 95 % confidence interval. The point reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the like-minded subsample (*Political Distance* 0-3) while the square and the corresponding line report the results for the contrary-minded subsample (*Political Distance* 4-7).

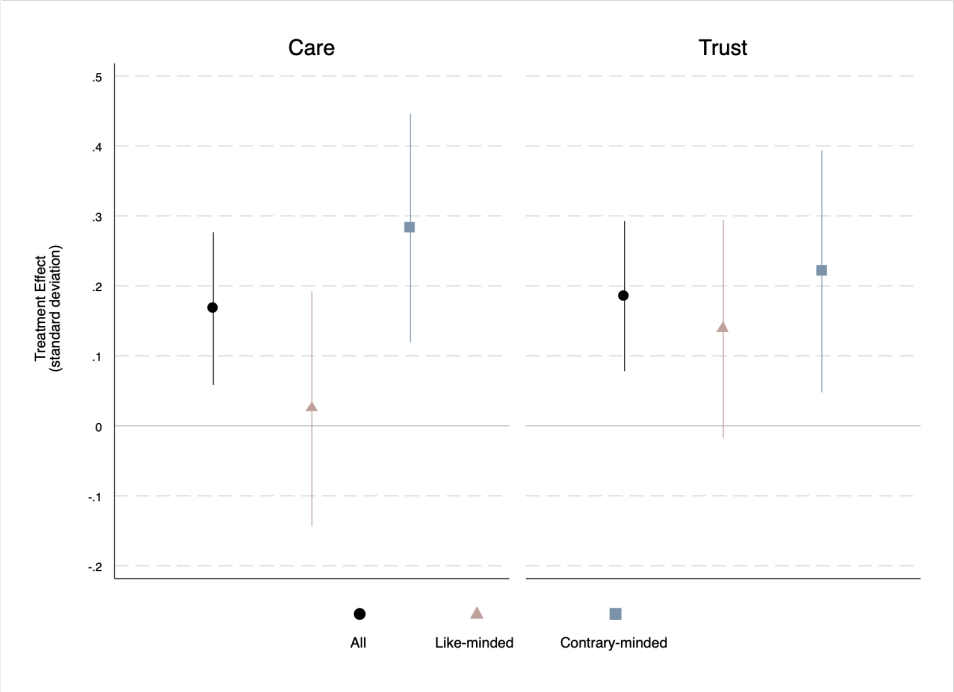
Figure 1.2. Effect on Willingness to Interact

1.4.2 Effect on Social Cohesion

On fear associated with the rising levels of affective polarization and growing heterogeneous beliefs between groups is the erosion of social cohesion. We study whether a face-to-face conversation could remedy this by looking at two important determinants for a cohesive society, namely how much citizens trust and care for each other. Analogue to before, meetings with *contrary-minded* partners should lead to greater effects. The reduction of stereotypical thinking and the reassessment of existing beliefs about other groups in society through intergroup contact could lead to a more positive assessment of society as a whole. Meeting a person who is alike should not affect the measures with the same magnitude, as individuals most likely meet a person from a familiar group.

Social Cohesion: Trust and Care. Figure 1.3 shows the effects on the two perceived social cohesion measures for the three samples. For the whole sample we observe that the discussion led to a significant change of 0.168 (*Care*) and 0.185 (*Trust*) standard deviations. Confirming our hypothesis, we observe that the mag-

nitude of the effects differ between the two subsamples. If someone met a partner with different political views we find that the conversation significantly bolstered perceived social cohesion. Participants that had a discussion with a *contrary-minded* partner increase their trust in other citizens by 0.221 standard deviations and also perceive Germans as more caring (0.283 standard deviations). For participants that met a *like-minded* partner the discussion also had positive effects. However, the effects are not or only weakly significant and also smaller in magnitude (0.0243 standard deviations for *Care* and 0.139 standard deviation for *Trust*). Tables 1.C.4 and 1.C.5 show the corresponding values and the regression results of our specifications. The findings do not differ across specifications.



The figure shows the treatment effect on the (standardized) two measures of perceived social cohesion. *Care* reflects the participants' agreement with the statement that the people in Germany generally care about the wellbeing of others and *Trust* is the participants' agreement to the statement that people in Germany can generally be trusted. The figure shows the estimated treatment coefficients of our main specification and the respective 95 % confidence intervals. The points report the treatment coefficients and the lines the 95 % confidence interval for the whole sample. Analogously, the triangles and the attached lines report the results for the like-minded subsample (*Political Distance* 0-3) while the squares and the corresponding lines report the results for the contrary-minded subsample (*Political Distance* 4-7).

Figure 1.3. Effect on Perceived Social Cohesion

Result 2. *A face-to-face conversation about politics strengthened social cohesion by increasing perceived trustworthiness and perceived care about others' wellbeing. The effects are driven by meetings between contrary-minded individuals.*

In contrast to our findings for stereotypes, there is no sign that meeting a *like-minded* partner might worsen the situation. Nevertheless, we observe that only conversations with a *contrary minded* partner significantly boost perceived social cohesion. Contact with a person that holds substantially different views seems to make participants realize that despite their differences, they "are all in the same boat".

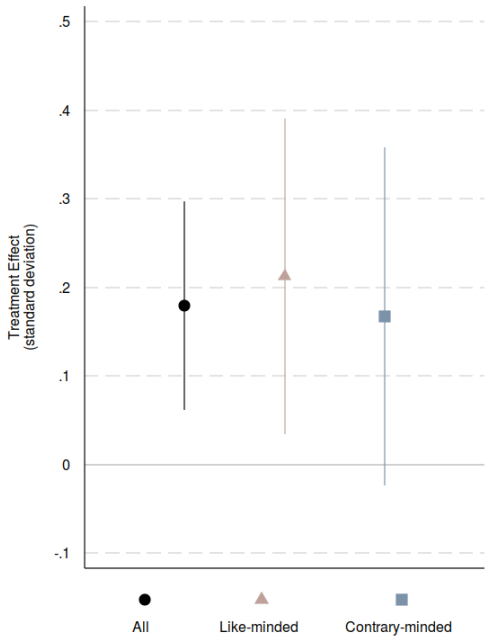
1.4.3 Effect on Political Attitudes

A natural next step is to investigate whether a political face-to-face discussion has the potential to shift political attitudes and fight issue polarization, particularly given the promising results above. In contrast to the previous subsections, the classic contact hypothesis plays a less important role. Instead, the exchange of information and perspectives during the meeting might challenge one's opinion on certain topics. As outlined in the introduction, there is research suggesting that the political orientation of the meeting partner matters. Having a discussion with *like-minded* partners may lead to a reinforcement of attitudes, while it is unclear whether we should expect a convergence of attitudes for *contrary-minded* pairs.

Total Adjustment: Challenging the Own Opinion

In a first step, we look at the *undirected movement* to answer the overarching question of whether the discussion pushed people to scrutinize, and ultimately change, their own viewpoints - independent of which direction the change occurred. Figure 1.4 shows the effect of the face-to-face conversation on standardized *PA_change* for the whole, the *like-minded* and the *contrary-minded* sample. Looking at all participants, we find that people change their political attitudes due to the meeting by 0.179 standard deviations.[27] This effect is driven by people who met *like-minded* partners changing their attitude by 0.212 standard deviations more than the corresponding people in the control group. The coefficient for *contrary-minded* pairs is positive as well, yet slightly smaller and only significant at 10%. Table 1.C.6 presents the exact values of the corresponding regressions and the results of the further specifications.

27. Note that we allow for change in the control group as well. We test whether participants in the treatment group changed their overall attitude significantly more than participants in the control group.



The figure shows the treatment effect on the (standardized) *PA_change*. It depicts the estimated treatment coefficient of our main specification and the respective 95 % confidence interval. The point reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the like-minded subsample (*Political Distance* 0-3) while the square and the corresponding line report the results for the contrary-minded subsample (*Political Distance* 4-7).

Figure 1.4. Effect on Attitudes: Undirected Adjustment

Result 3. *A political face-to-face conversation led to an adjustment of political attitudes. The effect is driven by participants who met a like-minded partner.*

The findings confirm the underlying idea of *Deutschland spricht*: A face-to-face discussion leads to questioning and adjusting one's opinion. This is a necessary condition for our next step in the analysis, an investigation of the direction of the adjustment.

Direction of Adjustment

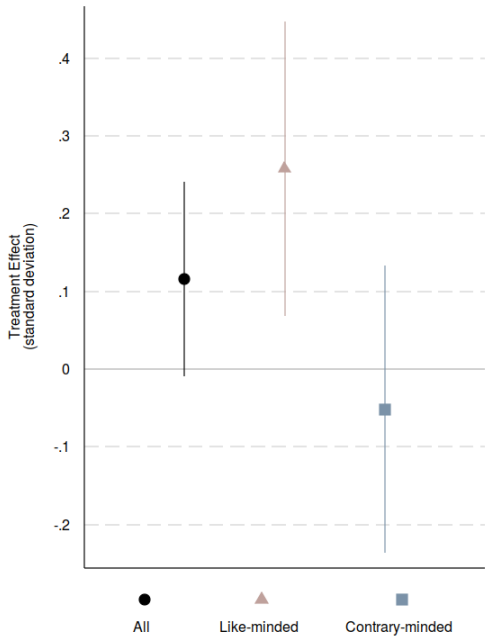
Next, we draw a more detailed picture of the adjustment by exploring its direction. First, we look at the movement towards the boundaries of our scale. Subsequently, we turn to the movement towards the average pre-meeting opinion.[28]

Adjustment towards Boundaries. Does a conversation lead to more moderate views or do individuals become more extreme in their opinions? Moreover, how does the direction of change depend on the partner's attitudes? Figure 1.5 answers these questions by presenting the impact of the face-to-face discussion on (standardized) *PA_change_center* for our three samples. As explained in detail in Section 1.2.2, the variable indicates whether a person moved away from ($PA\_change\_center > 0$) or closer to ($PA\_change\_center < 0$) the center of our seven point Likert scale.[29] Looking at the whole sample, we find a positive effect; however, it is only significant at a 10% level. We see that there is a pronounced adjustment for conversations between *like-minded* pairs that is significant at a 1% level: These participants moved 0.261 of a standard deviation towards the boundaries. In contrast, we do not observe a significant effect for participants that were matched with a partner that holds substantially different attitudes. Table 1.C.7 presents the coefficient estimates of all specifications confirming our findings.

If one is willing to assume that the movement towards the boundaries of the seven point Likert scales reflects a movement towards stronger or more extreme positions, the discussed results can be interpreted as follows. Discussions with *like-minded* partners reinforced political attitudes and intensified prior political positions. This result can arguably be interpreted as evidence that conversations with individuals that hold similar beliefs have the power to widen the trenches between different groups in a society as people move further apart from each other ideologically.

28. It would also be interesting to see whether people converged towards the attitude of their partner. Unfortunately, we do not have observations where we also know the partner's attitudes, which would be a necessary requirement for this analysis.

29. As before, we allow for movement in the control group as well. The treatment coefficient merely indicates whether the treatment group converged or diverged more than the control group.

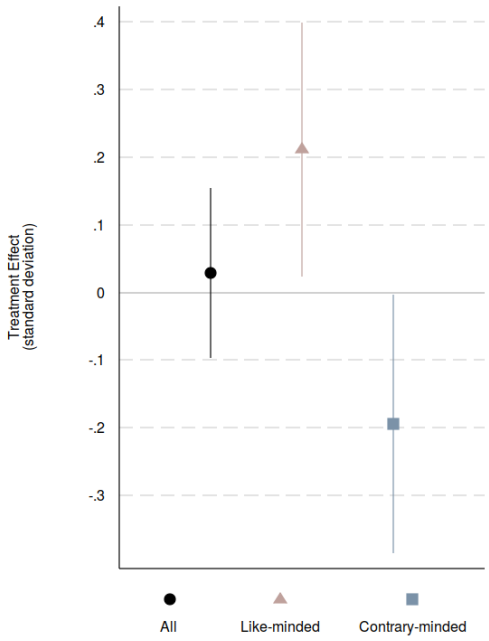


The figure shows the treatment effect on (standardized) *PA_change_center*. It depicts the estimated treatment coefficient of our main specification and the respective 95 % confidence interval. The point reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the like-minded subsample (*Political Distance* 0-3), while the square and the corresponding line report the results for the contrary-minded subsample (*Political Distance* 4-7).

Figure 1.5. Effect on Attitudes: Adjustment towards Boundaries

Convergence towards Average Opinion. Do participants converge or diverge from the average opinion due to a face-to-face conversation? We address this question by looking at (standardized) *PA_change_average*, which indicates the magnitude of a person's movement away from ($PA\_change\_average > 0$) or towards ($PA\_change\_average < 0$) the average opinion of our impact sample before the discussion.[30] Figure 1.6 plots the effects for our three samples.

30. Analogue to *PA_change_center*, a positive treatment coefficient means that people in the treatment group diverged further from the pre-meeting average than people in the control group. See Section 1.2.2 for more details on the construction.



The figure shows the treatment effect on (standardized) *PA_change_average*. It depicts the estimated treatment coefficient of our main specification and the respective 95 % confidence interval. The point reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the like-minded subsample (*Political Distance* 0-3), while the square and the corresponding line report the results for the contrary-minded subsample (*Political Distance* 4-7).

Figure 1.6. Effect on Attitudes: Convergence towards Average Opinion

We find no effect for the whole sample. The estimates for the *like-minded* and *contrary-minded* samples differ notably. Both significant at a 5% level, the coefficients are 0.213 and -0.193 standard deviations, respectively. Thus, the pattern is in line with our previous findings: Meeting a *like-minded* person drove the attitudes away from the average opinion, while conversations between *contrary-minded* individuals led to a convergence towards the average opinion. Table 1.C.8 shows all regression results.

Result 4. *If people met a like-minded partner, the face-to-face conversation moved their attitudes towards the boundaries of the scale and away from the pre-meeting average opinion of the sample. While meeting a contrary-minded partner did not drive attitudes towards or away from the boundaries, it did move participants closer to the average opinion.*

The results in this subsection reveal an interesting pattern: Meeting a person with similar views leads to an adjustment towards stronger opinions.[31] Thus, existing political views are reinforced by the conversation. On the other hand, for individuals that met a *contrary-minded* person we only observe an adjustment towards the average opinion. This could be seen as suggestive evidence that conversations with individuals who have different attitudes help to find common ground.

1.4.4 Robustness

In this subsection we further assess the robustness and external validity of our results. First, we present evidence to reduce the concern that participants' acceptance of the partner systematically depended on the political distance between the pair, which would affect our heterogeneous treatment effects. Second, we address selective attrition. Subsequently, we assess how well our intent-to-treat estimates capture the actual effects of the meeting. Last, we discuss a potential alternative interpretation of our results on stereotypes and social cohesion.

If the likelihood to accept the match first, i.e. the likelihood to enter the impact sample, depended on whether a person was matched to a contrary- or a like-minded person, the two subsamples might systematically consist of different types of persons. To rule out that the probability to accept first depends on whether a person was matched with a contrary- or a like-minded partner, we run several regressions using an *accept first* dummy as the dependent and a *contrary-minded* dummy an independent variable.[32] As shown in Table 1.D.1, we do not find that the probability of accepting the partner first depends on whether a person was assigned to a contrary- or a like-minded partner.

To assess the potential threat due to selective attrition, we estimate attrition rates for our treatment and control group. We find that attrition rates between the pre- and post-survey are very similar (49.9% and 47.6%, respectively). As many participants already knew their treatment status when the pre-survey was sent out[33], we also report how many first mover filled out both questionnaires depending on whether their partner accepted them (treatment group) or not (control group). However, besides attrition these rates this also capture the general willingness to fill out both questionnaires. The rates differ by 6.5 percentage points (16.6% and 23.5%).

31. As discussed in the respective paragraph on adjustment towards boundaries, we highlight that this interpretation is only valid if one is willing to interpret the scale in a way that its boundaries denote stronger or more extreme positions.

32. The *accept first* dummy is zero if a participant either did not accept the suggested partner or only accepted after the *first mover* accepted.

33. If the participant and her partner had already mutually accepted the match, treatment status was already known when the pre-survey was sent out. However, there was still uncertainty for the other cases, i.e. the case where nobody of the pair accepted yet and the case where the second mover did not yet accept. See Section 1.2 for details.

To get a feeling how well our intent-to-treat effect captures the real effect of a face-to-face meeting, we look at compliance with treatment assignment. In other words, we look at how many participants in our treatment group actually had a conversation.[34] We are able to quantify this as we asked participants in the post-survey whether they had met their partner or not. Among the 969 participants assigned to the treatment group, 87% claimed to have met their partner. Thus, once both partners agreed to have a discussion, it was very likely that the meeting actually took place. Overall, with compliance rates of 100% for the control group and 87% for the treatment group, we are confident that we capture the real effect of a meeting quite well.[35]

One possible argument against our interpretation of the results on stereotypes and social cohesion is that answers from participants in our control group might partly be driven by the disappointment of not being accepted by the suggested partner. To see if disappointment was a determining factor, we use data from people who were never rejected by the partner but nevertheless had no conversation.[36] We present evidence that there are no signs of disappointment in our control group by comparing the time-trend of the outcome variables between subjects who were never rejected and our control group.[37] Table 1.D.2 presents the diff-in-diff estimates. Confirming our interpretation, there is no sign of different time trends between the two groups.

1.5 Conclusion

This study explores the impact of political face-to-face conversations and provides important insights about their potential to fight growing social segregation in democracies. Our first main takeaway is that the effects of a conversation depend on the political opinion of the partner. When an individual had a conversation with a partner that holds substantially different political attitudes, stereotypes were reduced and perceived social cohesion improved. Furthermore, individuals slightly adapt their political attitudes and move towards the pre-conversation average opinion of the

34. As contact details were only exchanged after the second mover had accepted, it is impossible that participants were assigned to the control group but were in reality in the treatment group. There were two participants who stated that they met a partner even though the partner did not accept them. We dropped them from our main analysis, as we do not know whether they accidentally stated that they met, they lied on purpose, or they were able to meet without mutual acceptance. However, including them in our analysis does not change our results.

35. One could conjecture that we measure a lower bound of the real effect. There are people in our treatment group who did not meet anyone and it seems rather unlikely that these persons are the driving factor behind the positive effects on stereotypes and social cohesion.

36. These are participants who answered both surveys but never accepted their partner or were not matched.

37. Note that this comparison also makes use of the pre-survey data, which we carefully avoided in our main analysis. However, this might be less of a problem for the control group (see Section 1.3 for details).

sample. In contrast, conversations between like-minded individuals affected neither stereotypes nor perceived social cohesion significantly. If anything, we observe a negative tendency for stereotypes. When it comes to political attitudes, we see that these participants move towards the boundaries of the Likert scale and away from the average opinion of the whole sample. Both can be interpreted as a reinforcement of existing political beliefs.

Taken together, the results reveal a clear pattern. Meeting a person with sufficiently different views can have a positive impact and help to overcome the negative consequences of rising polarization. At the same time, staying within the own echo chamber and solely interacting with like-minded persons might exacerbate polarization and segregation of the society.

This paper should be seen as a proof of concept that, given the right circumstances, talking with others can indeed change a lot. However, there are some limitations to our study. Given that the people who participated in *Deutschland spricht* are not representative in terms of education, political orientation nor in their willingness to be confronted with contrasting views, it would be interesting to extend the analysis to a broader and more heterogeneous set of people. Moreover, future research could ascertain how long-lasting the effects are and how they transfer to real world behavior.

Appendix 1.A The Intervention

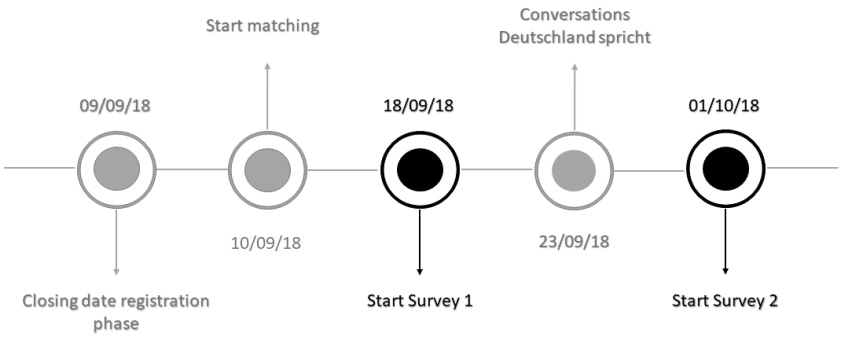


Figure 1.A.1. Timeline

Table 1.A.1. Political Registration Questions

Question	Answer Scale
Should Germany implement stronger border controls?	YES or NO
Did the public debate about sexual harassment and metoo caused something positive?	YES or NO
Should meat be taxed more to reduce the consumption of it?	YES or NO
Should German inner-cities be car-free?	YES or NO
Do Muslims and Non-Muslims cohabit well in Germany?	YES or NO
Are Germans worse off than 10 years ago?	YES or NO
Is Donald Trump good for the USA?	YES or NO

The table lists all seven political registration questions elicited during registration.

Table 1.A.2. Five Open Questions

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

Five open questions elicited during registration for *Deutschland Spricht*.

Appendix 1.B Descriptive Statistics

Table 1.B.1. Summary Statistics

	Impact Sample	
	mean	sd
Age	47.43	15.31
Female	0.37	0.48
Political Distance	3.48	1.89
Migration background	0.10	0.30
Political spectrum left-right	2.15	1.12
Political Registration Questions		
Border Control	0.27	0.45
Metoo	0.75	0.43
Meat Tax	0.63	0.48
Car-free inner-cities	0.61	0.49
Cohabitation Muslims	0.86	0.35
Development Germany	0.17	0.37
Trump	0.11	0.31
Education		
No Education	0.00	0.00
Lower Sec. Education	0.01	0.11
Middle School	0.07	0.25
Advanced technical certificate	0.06	0.24
High school	0.17	0.38
University	0.57	0.50
Doctorate	0.10	0.30
Other	0.01	0.12
Income (EUR)		
0-800	0.10	0.29
800-1500	0.13	0.34
1500-2200	0.20	0.40
2200-3300	0.23	0.42
3300+	0.27	0.44
Don't know	0.01	0.07
Party		
Die Linke	0.13	0.34
Bündnis/90 Die Grüne	0.47	0.50
SPD	0.10	0.30
FDP	0.07	0.26
CDU	0.06	0.24
AfD	0.06	0.24
Other	0.04	0.20
Don't Vote	0.02	0.12
Not specified	0.05	0.21
Don't know	0.00	0.00
Pol. het. of social environment		
No One	0.01	0.10
Nearly No One	0.10	0.31
Few	0.49	0.50
Ca. Half	0.25	0.44
Many	0.11	0.32
Nearly All	0.01	0.11
All	0.00	0.04
Religious confession		
Observations	1523	

Table 1.B.1. (continued)

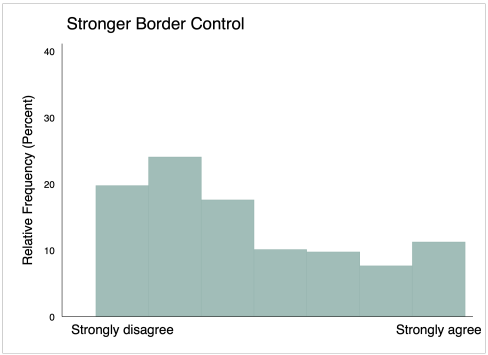
	Impact Sample	
	mean	sd
None	0.55	0.50
Christian	0.42	0.49
Islam	0.00	0.05
Buddhism	0.01	0.09
Jewish	0.00	0.06
Hindu	0.00	0.00
Other	0.00	0.05
Attendance house of worship		
Never	0.37	0.48
Rarely	0.35	0.48
Several Year	0.17	0.38
Several Month	0.05	0.22
Weekly	0.02	0.16
Several Week	0.02	0.13
Observations	1523	

The table reports descriptives for the sample we use in our analysis, the *first movers* who filled out both surveys.

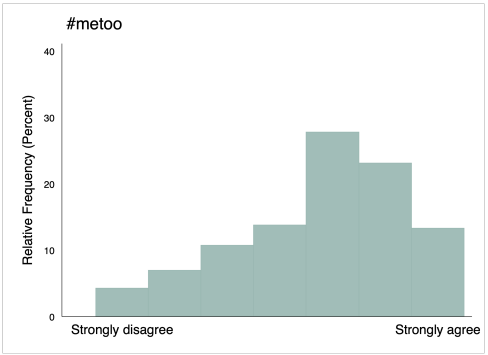
Table 1.B.2. Balance Checks

	All	Like-minded	Contrary-minded
Border Control	0.0120	0.103	-0.0813
	(0.0656)	(0.105)	(0.0954)
Metoo	-0.0540	-0.0785	-0.116
	(0.0666)	(0.0948)	(0.101)
Meat Tax	-0.126	-0.0245	-0.183
	(0.0791)	(0.108)	(0.126)
Car free inner-cities	-0.158**	-0.118	-0.143
	(0.0722)	(0.102)	(0.114)
Cohabitation (Non-)Muslims	0.0599	0.0329	0.0935
	(0.0637)	(0.0898)	(0.102)
Development Germany	0.0645	0.0455	0.0943
	(0.0756)	(0.117)	(0.114)
Trump	0.0532	0.0364	0.14
	(0.0551)	(0.0793)	(0.0885)
Importance: Border Control	0.0960	0.105	0.00937
	(0.101)	(0.165)	(0.142)
Importance: Metoo	-0.0497	0.00676	-0.172
	(0.0927)	(0.141)	(0.142)
Importance: Meat Tax	0.0673	0.118	0.0737
	(0.089)	(0.135)	(0.13)
Importance: Car free inner-cities	0.131	0.166	0.156
	(0.0887)	(0.127)	(0.137)
Importance: Cohabitation (Non-)Muslims	0.172**	0.201*	0.188
	(0.799)	(0.114)	(0.128)
Importance: Development Germany	0.118	0.186	0.11
	(0.102)	(0.158)	(0.149)
Importance: Trump	0.110	0.211	0.1
	(0.108)	(0.166)	(0.159)
Perspective taking I	-0.0712	-0.0825	-0.0122
	(0.0712)	(0.116)	(0.106)
Perspective taking II	0.0561	0.107	0.0324
	(0.0643)	(0.0917)	(0.104)
Social acceptance I	0.0401	-0.0087	0.0563
	(0.0632)	(0.0945)	(0.0985)
Social acceptance II	0.0442	0.0366	0.00442
	(0.0558)	(0.0821)	(0.0860)

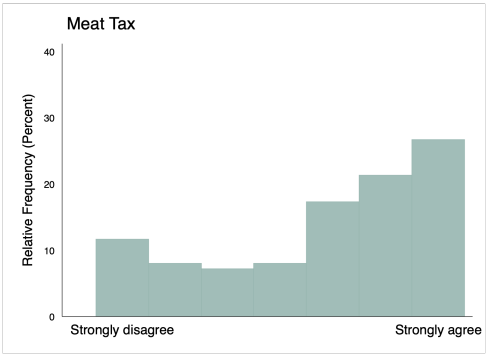
The table reports the treatment coefficients (*Treat*) of our balance checks. The dependent variables are 18 control and outcome variables from the pre-survey that were not used in our identification. The respective dependent variable is listed in the first column. Column (1) reports the results for the whole sample and column (2) and (3) for the two subsamples. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$



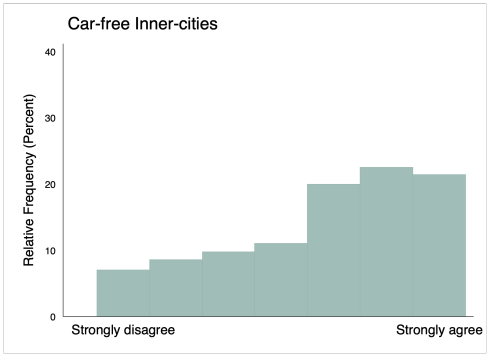
(a) Germany should implement stronger border controls.



(b) The public debate about sexual harassment and metoo caused something positive.



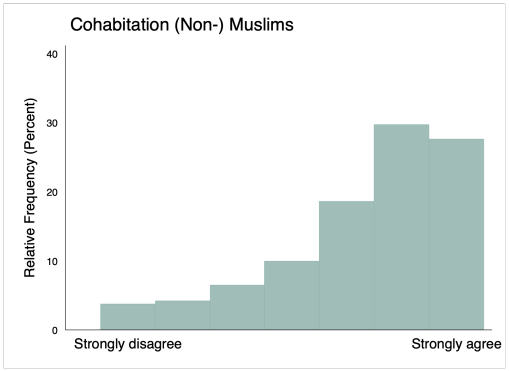
(c) Meat should be taxed more to reduce the consumption of it.



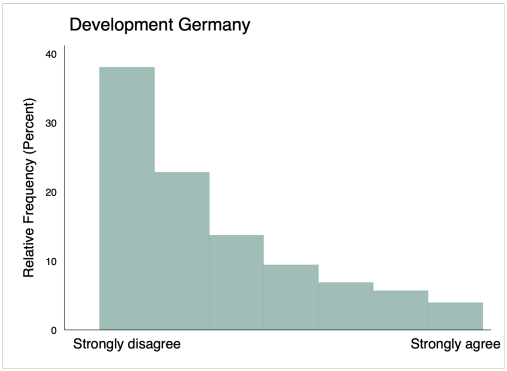
(d) German inner-cities should be car-free

The histograms plot the answers for four of the seven political attitude questions from the Pre-Survey. Participants had to state how much they agree to the respective statement (*0 = Strongly disagree* to *6 = Strongly agree*).

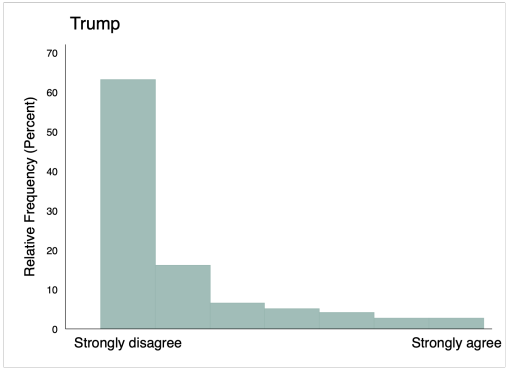
Figure 1.B.1. Answer Distributions Attitudes I



(a) Muslims and Non-Muslims cohabit well in Germany.



(b) Germans are worse off than 10 years ago.



(c) Donald Trump is good for the USA.

The histograms plot the answers for four of the seven political attitude questions from the Pre-Survey. Participants had to state how much they agree to the respective statement (*0 = Strongly disagree* to *6 = Strongly agree*).

Figure 1.B.2. Answer Distributions Attitudes II

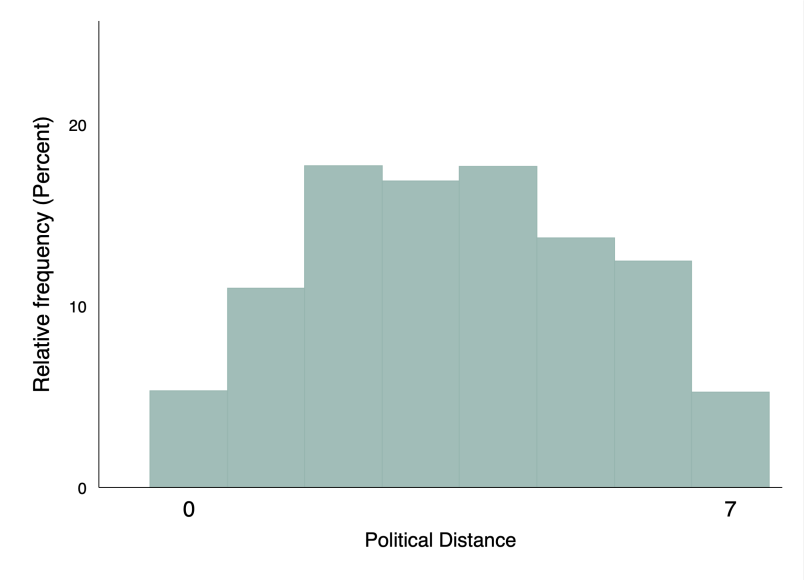


Figure 1.B.3. Political Distance within Pairs

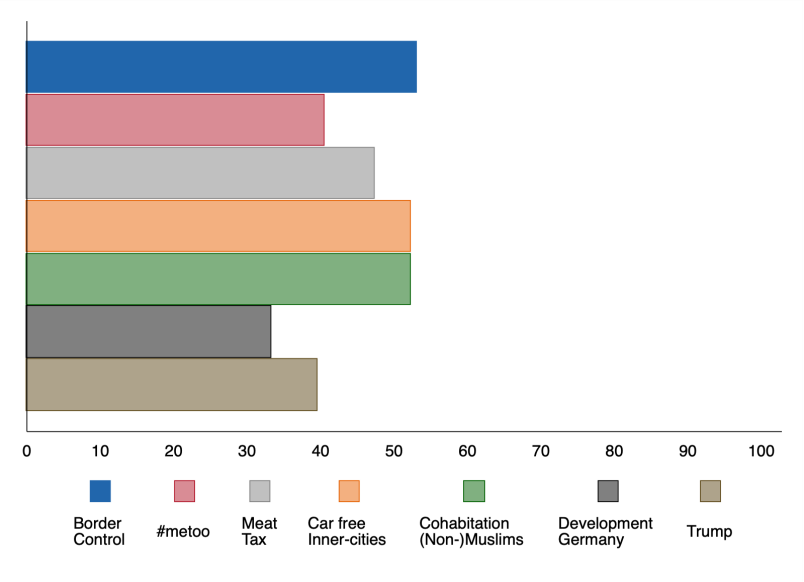


Figure 1.B.4. Topics during Conversation

Appendix 1.C Treatment Effects: Tables

Table 1.C.1. Effect on Stereotypes: Incompetence

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.106*	-0.102*	-0.0863	0.0416	0.0642	0.0536	0.0506	-0.329***	-0.319***	-0.293***
	(0.0567)	(0.0565)	(0.0560)	(0.0765)	(0.0833)	(0.0830)	(0.0809)	(0.0892)	(0.0872)	(0.0857)
Contrary-minded				0.0309						
				(0.0917)						
Treat × Contrary-minded				-0.316***						
				(0.113)						
Constant	-1.277***	-1.098***	-0.471**	-1.279***	-1.745***	-1.433***	-1.101***	-1.994**	-1.123*	-0.221
	(0.466)	(0.357)	(0.237)	(0.461)	(0.602)	(0.541)	(0.398)	(0.936)	(0.585)	(0.347)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.171	0.142	0.112	0.181	0.265	0.215	0.175	0.288	0.242	0.185
Observations	1474	1474	1474	1474	750	750	750	724	724	724

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about the *Incompetence* of people with very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.2. Effect on Stereotypes: Otherness

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.0336	-0.0211	0.0124	0.112	0.110	0.156*	0.177**	-0.201**	-0.200**	-0.156*
	(0.0607)	(0.0598)	(0.0595)	(0.0807)	(0.0862)	(0.0867)	(0.0833)	(0.0981)	(0.0956)	(0.0942)
Contrary-minded				0.120						
				(0.0939)						
Treat × Contrary-minded				-0.299**						
				(0.118)						
Constant	-0.764	-1.407***	-0.785***	-0.804	-1.907**	-1.718***	-1.245**	0.500	-1.174	-0.740**
	(0.559)	(0.428)	(0.275)	(0.558)	(0.775)	(0.640)	(0.494)	(1.021)	(0.764)	(0.373)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.155	0.130	0.0891	0.160	0.252	0.204	0.152	0.263	0.224	0.162
Observations	1475	1475	1475	1475	752	752	752	723	723	723

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about the *Otherness* of people with very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.3. Effect on Willingness to Interact

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.0283	-0.0214	-0.0211	-0.134*	-0.115	-0.127	-0.120	0.102	0.130	0.130
	(0.0564)	(0.0555)	(0.0548)	(0.0774)	(0.0875)	(0.0855)	(0.0827)	(0.0864)	(0.0841)	(0.0833)
Contrary-minded				-0.0609						
				(0.0902)						
Treat × Contrary-minded				0.219**						
				(0.109)						
Constant	-0.0402	0.886***	0.422*	-0.0212	-0.0227	0.939*	0.543	-0.330	0.866*	0.379
	(0.401)	(0.341)	(0.226)	(0.404)	(0.615)	(0.565)	(0.418)	(0.712)	(0.522)	(0.324)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.221	0.173	0.144	0.225	0.271	0.198	0.153	0.295	0.236	0.204
Observations	1482	1482	1482	1482	755	755	755	727	727	727

This table reports OLS estimates of the treatment effect on participant's willingness to interact with people who have very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample.

Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.4. Effect on Social Cohesion: Care

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.168***	0.160***	0.149***	0.0683	0.0243	0.000500	0.0137	0.283***	0.294***	0.252***
	(0.0556)	(0.0549)	(0.0552)	(0.0771)	(0.0856)	(0.0840)	(0.0822)	(0.0830)	(0.0824)	(0.0832)
Contrary-minded				-0.205**						
				(0.0865)						
Treat × Contrary-minded				0.187*						
				(0.107)						
Constant	-0.345	-0.371	-0.416*	-0.251	-0.226	-0.549	-0.233	-0.170	-0.123	-0.382
	(0.475)	(0.383)	(0.248)	(0.475)	(0.717)	(0.618)	(0.416)	(0.828)	(0.591)	(0.371)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.221	0.194	0.157	0.224	0.231	0.202	0.155	0.354	0.307	0.237
Observations	1486	1486	1486	1486	759	759	759	727	727	727

This table reports OLS estimates of the treatment effect on participant's standardized care with people that hold very different political attitudes on the treatment for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (Columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample.

Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.5. Effect on Social Cohesion: Trust

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.185***	0.177***	0.167***	0.185**	0.139*	0.122	0.125	0.221**	0.207**	0.176**
	(0.0547)	(0.0546)	(0.0541)	(0.0744)	(0.0793)	(0.0774)	(0.0760)	(0.0880)	(0.0886)	(0.0858)
Contrary-minded				-0.00592						
				(0.0846)						
Treat × Contrary-minded				0.000630						
				(0.109)						
Constant	-0.575	-0.0884	-0.269	-0.572	-0.531	-0.00873	0.00792	-0.585	0.0607	-0.124
	(0.575)	(0.373)	(0.217)	(0.578)	(0.796)	(0.588)	(0.379)	(0.851)	(0.552)	(0.285)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.277	0.254	0.221	0.277	0.250	0.206	0.143	0.409	0.371	0.327
Observations	1483	1483	1483	1483	757	757	757	726	726	726

This table reports OLS estimates of the treatment effect on participant's standardized trust with people that hold very different political attitudes on the treatment for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.6. Effect on Attitudes: Undirected Adjustment

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.179***	0.205***	0.182***	0.205**	0.212**	0.257***	0.236***	0.166*	0.168*	0.142
	(0.0602)	(0.0599)	(0.0593)	(0.0815)	(0.0909)	(0.0875)	(0.0852)	(0.0970)	(0.0975)	(0.0966)
Contrary-minded				0.00650						
				(0.0900)						
Treat × Contrary-minded				-0.0554						
				(0.118)						
Constant	1.074*	0.491	0.0162	1.077*	0.926	0.658	0.202	1.234	0.524	-0.00515
	(0.567)	(0.404)	(0.234)	(0.567)	(0.700)	(0.546)	(0.368)	(1.276)	(0.638)	(0.343)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.205	0.171	0.144	0.205	0.270	0.234	0.194	0.308	0.251	0.214
N	1416	1416	1416	1416	721	721	721	695	695	695

This table reports OLS estimates of the treatment effect on participant's standardized undirected adjustment *PA_change* with people that hold very different political attitudes on the treatment for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of *PA_change*.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.7. Effect on Attitudes: Adjustment towards the Boundaries

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.119*	0.109*	0.110*	0.252***	0.261***	0.255***	0.236***	-0.0465	-0.0545	-0.0483
	(0.0639)	(0.0634)	(0.0623)	(0.0886)	(0.0968)	(0.0963)	(0.0913)	(0.0940)	(0.0940)	(0.0909)
Contrary-minded				0.127						
				(0.0978)						
Treat × Contrary-minded				-0.268**						
				(0.122)						
Constant	-0.282	-0.566	-0.412	-0.327	-0.0677	-0.299	-0.538	-0.966	-1.231	-0.393
	(0.485)	(0.445)	(0.257)	(0.490)	(0.712)	(0.587)	(0.451)	(0.956)	(0.749)	(0.356)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.141	0.126	0.102	0.144	0.273	0.230	0.192	0.248	0.209	0.175
N	1416	1416	1416	1416	721	721	721	695	695	695

This table reports OLS estimates of the treatment effect on participant's standardized adjustment away from the center of the seven point Likert Scale *PA_change_center* with people that hold very different political attitudes on the treatment for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (Column 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Column 2,6 and 9 report regression results when controls for the open questions are added, column 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of *PA_change_center*. Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.C.8. Effect on Attitudes: Convergence towards Average Opinion

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.0301	0.0151	0.0295	0.183**	0.213**	0.191**	0.188**	-0.193**	-0.199**	-0.178*
	(0.0640)	(0.0627)	(0.0614)	(0.0857)	(0.0955)	(0.0914)	(0.0899)	(0.0969)	(0.0949)	(0.0917)
Contrary-minded				0.215**						
				(0.0960)						
Treat × Contrary-minded				-0.299**						
				(0.121)						
Constant	-0.317	-0.497	-0.389	-0.409	-0.109	-0.185	-0.427	-0.258	-1.249	-0.423
	(0.592)	(0.471)	(0.272)	(0.593)	(0.750)	(0.597)	(0.443)	(1.146)	(0.765)	(0.354)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.147	0.133	0.111	0.151	0.240	0.215	0.176	0.272	0.243	0.202
N	1416	1416	1416	1416	721	721	721	695	695	695

This table reports OLS estimates of the treatment effect on participant's standardized adjustment away from the average pre-meeting opinion *PA_change_average* with people that hold very different political attitudes on the treatment for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of *PA_change_average*.

Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 1.D Robustness

Table 1.D.1. Political Distance Dependent Selection

	All	Panel	Panel
Contrary-minded	-0.00553	0.0157	0.0316
	(0.00721)	(0.0187)	(0.0221)
Constant	0.446***	0.633***	-0.353
	(0.00488)	(0.0131)	(0.512)
Basic Controls	No	No	Yes
Name & Pol. Controls	No	No	Yes
Open Q. Controls	No	No	Yes
R^2	0.0000307	0.000267	0.0649
Observations	19135	2646	2582

The table reports OLS estimates. Dependent variable is a dummy equaling one if a person accepted first and zero if she accepted not or second. *Contrary-minded* is 1 if the *Political Distance* is larger than 3 and zero otherwise. The first columns contains all available observations while in columns (2) and (3) the sample is restricted to people who answered both surveys. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.D.2. Disappointment

	Incompetence	Otherness	Willingness	Care	Trust
time	0.201**	0.0478	-0.232**	0.0429	0.264***
	(0.0821)	(0.0709)	(0.0969)	(0.0855)	(0.0744)
control_alt	0.173	0.359***	-0.418**	-0.262*	-0.383**
	(0.136)	(0.121)	(0.172)	(0.155)	(0.153)
time × control_alt	-0.0156	-0.101	-0.0248	-0.0166	0.165
	(0.193)	(0.166)	(0.234)	(0.219)	(0.204)
Constant	-0.201***	-0.0439	3.569***	3.279***	3.953***
	(0.0596)	(0.0513)	(0.0703)	(0.0626)	(0.0561)
R^2	0.00752	0.0103	0.0154	0.00522	0.0214
Observations	1319	1321	1324	1328	1325

The table presents the results of regressions of our outcome variables (not standardized) on the dummy *time*, the dummy *control_alt* and their interaction. *time* equals zero before and one after the meeting. *control_alt* denotes whether a person belongs to the regular or the alternative control group which consists of those subjects in the panel who did not accept their partner. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 1.E Treatment Effects: Tables (Seperate Stereotypes)

Table 1.E.1. Effect on separate Stereotype: Moral Values

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.0611	-0.0603	-0.0371	0.104	0.120	0.143*	0.142*	-0.262***	-0.255***	-0.214**
	(0.0593)	(0.0584)	(0.0574)	(0.0798)	(0.0871)	(0.0854)	(0.0805)	(0.0960)	(0.0932)	(0.0901)
Contrary-minded				0.0845						
				(0.0931)						
Treat × Contrary-minded				-0.347***						
				(0.117)						
Constant	-0.209	-0.751*	-0.473*	-0.235	-1.215*	-1.064	-1.161**	0.338	-0.665	-0.238
	(0.508)	(0.412)	(0.265)	(0.509)	(0.728)	(0.656)	(0.493)	(0.981)	(0.638)	(0.341)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.131	0.114	0.0807	0.139	0.223	0.188	0.158	0.240	0.203	0.143
N	1477	1477	1477	1477	753	753	753	724	724	724

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about the *moral values* of people with very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.E.2. Effect on separate Stereotype: Way of Life

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	0.0135	0.0268	0.0610	0.0887	0.0813	0.114	0.147*	-0.0577	-0.0633	-0.0273
	(0.0583)	(0.0578)	(0.0576)	(0.0794)	(0.0837)	(0.0842)	(0.0819)	(0.0903)	(0.0892)	(0.0893)
Contrary-minded				0.107						
				(0.0919)						
Treat × Contrary-minded				-0.148						
				(0.114)						
Constant	-1.075**	-1.599***	-0.772***	-1.119**	-1.774**	-1.683***	-0.827*	0.281	-1.465*	-0.942***
	(0.539)	(0.420)	(0.250)	(0.537)	(0.750)	(0.620)	(0.478)	(0.916)	(0.759)	(0.329)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.154	0.133	0.0938	0.155	0.238	0.204	0.145	0.283	0.240	0.183
N	1479	1479	1479	1479	755	755	755	724	724	724

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about the *way of life* of people with very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.E.3. Effect on separate Stereotype: Cognitive Abilities

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.132**	-0.137**	-0.127**	0.0344	0.0410	0.0164	0.0175	-0.345***	-0.344***	-0.332***
	(0.0577)	(0.0573)	(0.0568)	(0.0777)	(0.0845)	(0.0847)	(0.0817)	(0.0892)	(0.0874)	(0.0860)
Contrary-minded				0.0727						
				(0.0919)						
Treat × Contrary-minded				-0.352***						
				(0.113)						
Constant	-1.125**	-0.967***	-0.432*	-1.153**	-1.889***	-1.430***	-0.979**	-1.165	-0.816	-0.163
	(0.473)	(0.352)	(0.235)	(0.470)	(0.616)	(0.532)	(0.399)	(1.012)	(0.595)	(0.344)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.157	0.134	0.108	0.167	0.242	0.194	0.161	0.271	0.233	0.185
N	1477	1477	1477	1477	753	753	753	724	724	724

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about the *cognitive abilities* of people with very different political attitudes for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 1.E.4. Effect on separate Stereotype: Badly Informed

	Whole Sample				Like-minded Sample			Contrary-minded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat	-0.0590	-0.0508	-0.0284	0.0509	0.0934	0.0900	0.0818	-0.260***	-0.240***	-0.204**
	(0.0566)	(0.0564)	(0.0558)	(0.0776)	(0.0835)	(0.0833)	(0.0819)	(0.0873)	(0.0853)	(0.0844)
Contrary-minded				-0.00497						
				(0.0924)						
Treat × Contrary-minded				-0.239**						
				(0.113)						
Constant	-1.162**	-1.010***	-0.465*	-1.147**	-1.391**	-1.262**	-1.113***	-2.390***	-1.130*	-0.261
	(0.484)	(0.369)	(0.237)	(0.481)	(0.626)	(0.566)	(0.413)	(0.880)	(0.582)	(0.339)
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name & Pol. Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Open Q. Controls	Yes	No	No	Yes	Yes	No	No	Yes	No	No
R^2	0.161	0.133	0.102	0.168	0.270	0.218	0.170	0.294	0.245	0.174
N	1478	1478	1478	1478	753	753	753	725	725	725

This table reports OLS estimates of the treatment effect on participant's stereotypical belief about how *badly informed* people with very different political attitudes are for the whole sample (columns 1 to 4), the like-minded subsample with *Political Distance* lower or equal than 3 (columns 5 to 7) and the contrary-minded subsample with *Political Distance* larger than 3 (columns 8 to 10). Columns 1,5 and 8 report the regression results using our main identification. Columns 2,6 and 9 report regression results when controls for the open questions are added, columns 3,7 and 10 depict the results if only basic controls are used. Column 4 reports the regression results using all controls plus the interaction effect with $Contrary - minded$, a dummy that is 1 if the individual has a *Political Distance* larger than 3, i.e. is in the contrary-minded subsample. See section 1.2.2 for more details on the construction of the dependent variable.
Robust Standard errors in parentheses. Basic controls are dummies for age intervals, gender, 2 digit zip-code and seven political registration questions. Name & Pol. Controls are dummies for education, income, migration background, political parties and political self-classification. Open Q. Controls are dummies for religion, religiousness, marital status and number of politically contrary-minded person in social environment. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 1.F Surveys

Note: The Instructions were translated into English by the authors.

1.F.1 Pre-Survey

Thank you for your participation. Please answer the following questions. If you are unsure about an answer to a question, simply choose the answer that you agree with the most.

1. 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
2. Were you and both your parents born in Germany?
   - Yes
   - No
3. Many people use the words 'left' and 'right' to describe political convictions. Below you will find a scale that goes from 'left' to 'right'. When you think about your own political attitudes, where would you place yourself on the scale below?
   - Left O O O O O O O Right
4. If there was a federal election next Sunday, which party would you vote for?
   - CDU/CSU
   - SPD
   - FDP
   - Bündnis 90/Die Grüne
   - Die Linke
   - AfD
   - Different party
   - I would not vote
   - I don't know
   - Prefer not to say

5. Did you do one or more of the following things in the last year?
   - Participate in a citizens' initiative
   - Participate in a demonstration
   - Work for a political party
   - Work for a union
   - None of the above
   - Prefer not to say
6. How many people in your personal environment (e.g. friends, family members, colleagues) have compared to you very different political and social views?
   - No one
   - Nearly no one
   - Some
   - About half
   - Many
   - Nearly all
   - All
   - Prefer not to say
7. What is your denomination?
   - Christianity
   - Islam
   - Buddhist
   - Jewish
   - Hindu
   - Different denomination
   - Without denomination
   - No response
8. How often do you visit a house of worship?
   - More than once a week
   - Once a week
   - One to three times a month
   - A few times in a year
   - Once or twice a year
   - Never
   - Prefer not to say

To what extent do the following statements apply to you? Please give your answer on a scale from "Not at all" (0) to "Fully applies" (6).

9. I have the feeling that most people take my opinion serious.
   - Not at all O O O O O O O Fully applies
10. I feel accepted.
   - Not at all O O O O O O O Fully applies

The questions in the next part relate to *Deutschland spricht*.

11. Did you already participate last year?
   - Yes
   - No

What were you expectations when you signed up for *Deutschland spricht*? We listed two possible expectations:

12. I hope to convince my partner from my point of view.
   - Not at all O O O O O O O Fully applies
13. I hope to learn something about my partner's point of view.
   - Not at all O O O O O O O Fully applies

During your registration for *Deutschland spricht* you answered seven Yes or No questions. Now we would like to ask you for a more detailed assessment. Please state to each of the following seven statements on a Scale from "Do not agree at all"(0) to "Fully agree"(6) how much you agree.

14. Germany should implement stronger border controls.
   - Do not agree at all O O O O O O O Fully agree
15. The public debate about sexual harassment and metoo caused something positive.
   - Do not agree at all O O O O O O O Fully agree
16. Meat should be taxed more to reduce its consumption.
   - Do not agree at all O O O O O O O Fully agree
17. German inner-cities should be car-free.
   - Do not agree at all O O O O O O O Fully agree
18. Muslims and Non-Muslims cohabit well in Germany.
   - Do not agree at all O O O O O O O Fully agree
19. Germans are worse off than 10 years ago.

- Do not agree at all O O O O O O O Fully agree

20. Donald Trump is good for the USA.
    - Do not agree at all O O O O O O O Fully agree

In a next step, we ask you to state how important the respective topics are for you.

21. Germany should implement stronger border controls.
    - Not important at all O O O O O O O Very important
22. The public debate about sexual harassment and metoo caused something positive.
    - Not important at all O O O O O O O Very important
23. Meat should be taxed more to reduce its consumption.
    - Not important at all O O O O O O O Very important
24. German inner-cities should be car-free.
    - Not important at all O O O O O O O Very important
25. Muslims and Non-Muslims cohabit well in Germany.
    - Not important at all O O O O O O O Very important
26. Germans are worse off than 10 years ago.
    - Not important at all O O O O O O O Very important
27. Donald Trump is good for the USA.
    - Not important at all O O O O O O O Very important

Now please picture a person that, compared to you, answered the seven questions **very differently**, i.e. a person that gave other answers to the questions than you. What do you think, how much do the following statements about this person apply?

28. This person has completely different moral values.
    - Not at all O O O O O O O Fully applies
29. This person leads a completely different life.
    - Not at all O O O O O O O Fully applies
30. This person is badly informed.
    - Not at all O O O O O O O Fully applies
31. This person is incapable of understanding complex contexts.
    - Not at all O O O O O O O Fully applies
32. I don't want this person to be in my personal environment.
    - Not at all O O O O O O O Fully applies

Next, we ask you to picture an average Bündnis 90/Die Grüne voter. What do you think, how much do the following statements apply to an average Bündnis 90/Die Grüne voter?

33. German inner-cities should be car-free. An average Bündnis 90/Die Grüne voter...
    - Does not agree at all O O O O O O O Fully agrees
34. Meat should be taxed more to reduce consumption. An average Bündnis 90/Die Grüne voter...
    - Does not agree at all O O O O O O O Fully agrees

Next, we ask you to picture an average AfD voter. What do you think, how much do the following statements apply to an average AfD voter?

35. Muslims and Non-Muslims cohabit well in Germany. An average AfD voter...
    - Does not agree at all O O O O O O O Fully agrees
36. The public debate about sexual harassment and metoo caused something positive. An average AfD voter...
    - Does not agree at all O O O O O O O Fully agrees

Now we want you to make an estimate.

37. How many asylum applications did the German government receive in 2017?
38. What percentage of the German population are Muslims?

In the final step, we would like to know how much you agree with the following general statements.

39. Marriage should only be possible between a men and a women.
    - Do not agree at all O O O O O O O Fully agree
40. You can trust most people in Germany.
    - Do not agree at all O O O O O O O Fully agree
41. Most people in Germany do not care about the wellbeing of their fellow citizens.
    - Do not agree at all O O O O O O O Fully agree
42. Germany should deepen its cooperation with other EU countries.
    - Do not agree at all O O O O O O O Fully agree
43. To reduce the gap between rich and poor, the maximum tax rate for top earner should be increased.
    - Do not agree at all O O O O O O O Fully agree
44. Altogether the German media is trustworthy.
    - Do not agree at all O O O O O O O Fully agree

1.F.2 Post-Survey

Thank you for your participation. Please answer the following questions. If you are unsure about an answer to a question simply choose the answer that you agree with the most. If you already answered the first questionnaire, you might recognize some of the questions. Please also answer these questions based on your best conscience. You help us a lot if you answer all questions.

1. Did you meet the matched partner?
   - Yes
   - No
   - I did not get a match.

If [1.] == Yes

A.2 When did you meet your partner?

A.3 How long did you talk?

To what extent do the following statements apply to your conversation?

A.4 The atmosphere was enjoyable.
   - Not at all O O O O O O O Fully applies

A.5 The discussion was heated.
   - Not at all O O O O O O O Fully applies

A.6 My partner is likable.
   - Not at all O O O O O O O Fully applies

A.7 Our political attitudes converged.
   - Not at all O O O O O O O Fully applies

A.8 I was able to convince my partner about my view.
   - Not at all O O O O O O O Fully applies

A.9 I learned a lot about my partner's views.
   - Not at all O O O O O O O Fully applies

A.10 My partner was able to convince me about his/her view.
   - Not at all O O O O O O O Fully applies

A.11 What topics did you discuss?

If [1.] == No

B.2 Why did you not meet your partner?
   - I accepted the partner but the partner did not accept me.

- I did not accept the partner.
- We both accepted, but still did not meet.

B.3 Did you like the partner's behavior?

- Yes
- No

If [B.2] == I did not accept the partner.

B.4 What was the reason that you did not accept the partner?

- Our views were too similar.
- The partner's view were too extreme.
- My partner was not likable.
- I wouldn't felt comfortable meeting the partner.
- My partner was too inexperienced.
- I had an appointment.
- I lost interest.
- Different reason.

4. What is your family status?
   - Married and living together with partner
   - Married not living together
   - Civil union
   - Widowed
   - Divorced
   - In a relationship
   - Single
   - Prefer not to say
5. What is your monthly household net-income?
   - Below 1300 EUR
   - 1300 to 2600 EUR
   - 2600 to 3600 EUR
   - More than 3600
   - Prefer not to say

We now would like to know from you how the following statements apply to you. Please answer on a Scale from "Not at all"(0) to "Fully applies"(6).

6. I find it difficult to take the perspective of another person.

- Not at all O O O O O O O Fully applies

7. Before I choose on a point of view i try to consider different perspectives.
   - Not at all O O O O O O O Fully applies
8. I have the feeling that most people take my opinion serious.
   - Not at all O O O O O O O Fully applies
9. I feel accepted.
   - Not at all O O O O O O O Fully applies

We would like to know how much you agree with the following general statements. Please state to each of the following seven statements on a Scale from "Do not agree at all"(0) to "Fully agree"(6) how much you agree.

10. Marriage should only be possible between a men and a women.
    - Do not agree at all O O O O O O O Fully agree
11. "Deutschland spricht" can help to improve the social coexistence.
    - Do not agree at all O O O O O O O Fully agree
12. Most people in Germany do not care about the wellbeing of their fellow citizens.
    - Do not agree at all O O O O O O O Fully agree
13. Germany should deepen its cooperation with other EU countries.
    - Do not agree at all O O O O O O O Fully agree
14. To reduce the gap between rich and poor, the maximum tax rate for top earner should be increased.
    - Do not agree at all O O O O O O O Fully agree
15. Altogether the German media is trustworthy.
    - Do not agree at all O O O O O O O Fully agree
16. You can trust most people in Germany.
    - Do not agree at all O O O O O O O Fully agree

Now we would like you to make an estimate.

17. How many asylum applications did the German government receive in 2017?

17a. Did you talk about the number?

- Yes
- No
- I don't know

18. What percentage of the German population are Muslims?

18a. Did you talk about the number?

- Yes
- No
- I don't know

Now we are returning to the seven Yes or No questions you answered during the registration for *Deutschland spricht*. We would like to ask you for a more detailed assessment. Please state to each of the following seven statements on a Scale from "Do not agree at all"(0) to "Fully agree"(6) how much you agree.

19. Germany should implement stronger border controls.
    - Do not agree at all O O O O O O O Fully agree
20. The public debate about sexual harassment and metoo caused something positive.
    - Do not agree at all O O O O O O O Fully agree
21. Meat should be taxed more to reduce its consumption.
    - Do not agree at all O O O O O O O Fully agree
22. German inner-cities should be car-free.
    - Do not agree at all O O O O O O O Fully agree
23. Muslims and Non-Muslims cohabit well in Germany.
    - Do not agree at all O O O O O O O Fully agree
24. Germans are worse off than 10 years ago.
    - Do not agree at all O O O O O O O Fully agree
25. Donald Trump is good for the USA.
    - Do not agree at all O O O O O O O Fully agree

Now please picture a person that compared to you answered the seven questions **very differently**, i.e. a person that gave other answers to the questions than you. What do you think, how much do the following statements about this person apply?

26. This person has completely different moral values.
    - Not at all O O O O O O O Fully applies
27. This person leads a completely different life.
    - Not at all O O O O O O O Fully applies

28. I This person is badly informed.
    - Not at all O O O O O O O Fully applies
29. This person is incapable of understanding complex contexts.
    - Not at all O O O O O O O Fully applies

30. I don't want this person to be in my personal environment.
    - Not at all O O O O O O O Fully applies

Next, we ask you to picture an average Bündnis 90/Die Grüne voter. What do you think, how much do the following statements apply to an average Bündnis 90/Die Grüne voter?

31. German inner-cities should be car-free. An average Bündnis 90/Die Grüne voter...
    - Does not agree at all O O O O O O O Fully agrees
32. Meat should be taxed more to reduce consumption. An average Bündnis 90/Die Grüne voter...
    - Does not agree at all O O O O O O O Fully agrees

Next, we ask you to picture an average AfD voter. What do you think, how much do the following statements apply to an average AfD voter?

33. Muslims and Non-Muslims cohabit well in Germany. An average AfD voter...
    - Does not agree at all O O O O O O O Fully agrees
34. The public debate about sexual harassment and metoo caused something positive. An average AfD voter...
    - Does not agree at all O O O O O O O Fully agrees

References

Abramowitz, Alan I, and Kyle L Saunders. 2008. "Is polarization a myth?" *Journal of Politics* 70 (2): 542–555. [9]

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew. 1954. "The nature of prejudice." [6]

Becker, Gary S. 1958. "Competition and democracy." *Journal of Law and Economics* 1: 105–109. [5]

Bertrand, Marianne, and Emir Kamenica. 2018. "Coming apart? Cultural distances in the United States over time." Techreport. National Bureau of Economic Research. [9]

Bishop, Bill. 2009. *The big sort: Why the clustering of like-minded America is tearing us apart*. Houghton Mifflin Harcourt. [5]

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. [7]

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2019. "Cross-Country Trends in Affective Polarization." [5, 9]

Broockman, David, and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352 (6282): 220–224. [7]

Burns, Justine, Lucia Corno, and Eliana La Ferrara. 2015. "Interaction, prejudice and performance. Evidence from South Africa." *Unpublished working paper. https://www. povertyactionlab. org/sites/default/files/publications/5167_Interations% 2Cprejudiceand-performance_Eliana_March2015. pdf*, [7]

Center, Pew Research. 2014. "Political polarization in the American public." [9]

Chen, M Keith, and Ryne Rohla. 2018. "The effect of partisanship and political advertising on close family ties." *Science* 360 (6392): 1020–1024. [9]

DeMarzo, Peter M, Dimitri Vayanos, and Jeffrey Zwiebel. 2003. "Persuasion bias, social influence, and unidimensional opinions." *Quarterly journal of economics* 118 (3): 909–968. [8]

Desmet, Klaus, Ignacio Ortuño-Ortın, and Romain Wacziarg. 2017. "Culture, ethnicity, and diversity." *American Economic Review* 107 (9): 2479–2513. [9]

Dimock, Michael, Carroll Doherty, Jocelyn Kiley, and Russ Oates. 2014. "Political polarization in the American public." *Pew Research Center*, [5]

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *Journal of political economy* 65 (2): 135–150. [5]

Druckman, James N, Erik Peterson, and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American Political Science Review* 107 (1): 57–79. [9]

Enke, Benjamin. 2018. "Moral values and voting." Techreport. National Bureau of Economic Research. [9]

Enke, Benjamin, and Florian Zimmermann. 2017. "Correlation neglect in belief formation." *Review of Economic Studies* 86 (1): 313–332. [8]

Eyster, Erik, and Matthew Rabin. 2014. "Extensive imitation is irrational and harmful." *Quarterly Journal of Economics* 129 (4): 1861–1898. [8]

Finseraas, Henning, Torbjørn Hanson, Åshild A Johnsen, Andreas Kotsadam, and Gaute Torsvik. 2016. "Trust, ethnic diversity, and personal contact: Experimental field evidence." In *Seminar presentation at Universidad Carlos III*. [7]

Fiorina, Morris P, and Samuel J Abrams. 2008. "Political polarization in the American public." *Annu. Rev. Polit. Sci.* 11: 563–588. [9]

Flaxman, Seth, Sharad Goel, and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80 (S1): 298–320. [9]

Gentzkow, Matthew. 2016. "Polarization in 2016." Techreport. Stanford University. [5, 9]

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78 (1): 35–71. [9]

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

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87 (4): 1307–1340. [9]

Glaeser, Edward L, and Cass R Sunstein. 2009. "Extremism and social learning." *Journal of Legal Analysis* 1 (1): 263–324. [7]

Glaeser, Edward L, and Bryce A Ward. 2006. "Myths and realities of American political geography." *Journal of Economic Perspectives* 20 (2): 119–144. [9]

Gutmann, Amy, and Dennis F Thompson. 1998. *Democracy and disagreement*. Harvard University Press. [8]

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. [5, 9]

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, not ideologya social identity perspective on polarization." *Public opinion quarterly* 76 (3): 405–431. [9]

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. [5, 9]

Lelkes, Yphtach. 2016. "Mass polarization: Manifestations and measurements." *Public Opinion Quarterly* 80 (S1): 392–410. [9]

Munzert, Simon, and Paul C Bauer. 2013. "Political depolarization in German public opinion, 1980–2010." *Political Science Research and Methods* 1 (1): 67–89. [9]

Paluck, Elizabeth Levy. 2016. "How to overcome prejudice." *Science* 352 (6282): 147–147. [7, 8]

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. [7]

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review* 109 (3): 774–809. [7]

Schkade, David, Cass R Sunstein, and Reid Hastie. 2007. "What happened on deliberation day." *Cal. L. Rev.* 95: 915. [7]

Sunstein, Cass R. 1999. "The law of group polarization." *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91): [7]

Sunstein, Cass R. 2009. *Going to extremes: How like minds unite and divide*. Oxford University Press. [7]

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

Wojcieszak, Magdalena. 2011. "Deliberation and attitude polarization." *Journal of Communication* 61 (4): 596–617. [8]

Chapter 2

Stereotypes about Refugees - how motives mold peoples' stereotypes

2.1 Introduction

"[...] the actual foundation of racism is not ignorance and hate, but self-interest, particularly economic and political and cultural."
– Dr. Ibram X. Kendi

It is unfortunately not uncommon that our behavior towards others is incompatible with the idea that we are intrinsically good people. In such situations, our actions and decisions conflict with our desire to think of ourselves as altruistic, moral and generous.[1] How do we resolve this tension between selfish actions and the preservation of a positive self-image? According to studies, one way individuals self-exculpate is by distorting related beliefs and preferences, by, for example, modifying their beliefs about others' altruism or fairness, or their own risk and ambiguity preferences in order to cast their actions in a positive light (see, e.g. Konow (2000), Haisley and Weber (2010), Di Tella, Perez-Truglia, Babino, and Sigman (2015), and Exley (2016)).
This paper studies how far individuals are willing to go to rationalize their behavior, by exploring whether individuals adopt extremely negative beliefs or stereotypes about a minority social group to justify their own selfish behavior. To test this, our study focuses on an often marginalized group: refugees. In 2015 and 2016,

1. This desire for a positive self-image is a central assumption in a number of theoretical models (see, e.g. Bodner and Prelec (2003), Bénabou and Tirole (2006), Bénabou and Tirole (2011), and Bénabou, Falk, and Tirole (2019)) and has been shown to be a driving factor for human behavior in several experiments (see, e.g. Ariely, Bracha, and Meier (2009), Falk (2017), and Grossman and Van der Weele (2017))

over 2.5 million individuals applied for asylum in the EU. The unprecedented number of individuals seeking refuge gave rise to anti-immigration sentiments all over Europe.[2] The emergence of anti-immigration parties and sentiments was accompanied by a rise in stereotypical thinking about refugees. Alesina, Miano, and Stantcheva (2018) and Alesina, Murard, and Rapoport (2019) show that a large share of European citizens has vastly wrong beliefs about asylum seekers, with respondents of their large-scale, Europe-wide survey dramatically overestimating the number of individuals seeking asylum and the economic, cultural and religious distance between them and the refugees. But how do these large misperceptions arise? We argue that individuals have motives to engage in belief distortion, in particular characterized by a desire to justify their support for certain policies or for their selfish actions. To showcase how motives mold peoples' stereotypes, this paper seeks to address whether individuals, in situations in which they can benefit at the expense of refugees, distort their beliefs about refugees' treatment of women?

The basic idea underlying motivated reasoning is that the formation of beliefs is not always driven by the desire to be accurate. Instead, some beliefs are formed to protect other desired beliefs, e.g. thinking about oneself as moral or selfless. In situations in which an individual's actions conflict with the desire to think about oneself as selfless, the individual reacts by distorting certain beliefs to uphold a positive self-assessment. In the context of our study, we expect that individuals rationalize selfish actions towards refugees by adopting negative stereotypical views about them. To show this, we need to create a situation in which some individuals can act selfishly towards refugees while at the same time having the opportunity to state a belief about refugees' treatment of women. We construct an online survey experiment that explicitly creates such situations. By varying the possibility to enrich oneself at the expense of refugees, we can present causal evidence on motivated stereotypes. In total, 902 individuals participated in our online experiment. The sample is representative for the adult German population.[3] During the online experiment, the participants were randomly allocated to one of three experimental conditions, consisting of one treatment (*Main Treatment*) and two control conditions (*Control Self-interest* and *Control Context*). All three conditions had an identical structure. After filling out a short questionnaire and receiving all relevant information regarding their upcoming choices, participants had to make two decisions. The two decisions were displayed on a single page - we will call this page the 'decision page' from now on. Across all experimental

2. The most blatant example of these anti-immigration or anti-refugee sentiments is the surge of right-wing parties in elections throughout Europe (see, e.g. Barone, D'Ignazio, Blasio, and Naticchioni (2016), Halla, Wagner, and Zweimüller (2017), and Dustmann, Vasiljeva, and Piil Damm (2019)).

3. The sample is representative for the adult German population in the following key variables: age (divided in groups), income groups, gender, distribution of East and West German citizens and education (the percentage of people that hold a university degree).

conditions we elicited our main outcome variable participants' *misogyny belief*. The *misogyny belief* is the participant's incentivized guess of what percentage of refugees that arrived in Germany between 2013 and 2016 stated that women should in no case have equal rights in a democracy.[4] The second decision varied across the different conditions: In *Main Treatment* participants were able to take away money from a 50 EUR donation to a large and widely known pro-immigration organization to give to themselves.[5] For every 1 EUR taken away only 50 cents was given to the participants, making it relatively selfish to withdraw money. We hypothesize that participants, when taking money away from the donation, are motivated to distort their *misogyny belief*. In our first control (*Control Self-interest*) participants were able to allocate 50 EUR between two pro-immigration organizations.[6] By making it impossible to take away some additional money for themselves, we rule out that participants' answers to the *misogyny belief* question were driven by the motive to rationalize selfish behavior. In our third condition (*Control Context*) participants were able to take away money from a donation to an environmental organization. Again, for every 1 EUR taken the participant only received 50 cents. As in the other control condition, we rule out that participants' answers to the *misogyny belief* question were driven by the motive to rationalize their own behavior. We do so by taking away the connection between the donation decision and the *misogyny belief*.[7]

Comparing participants' *misogyny beliefs* between *Main Treatment* and each of the two controls separately produces causal evidence of whether individuals instrumentalize negative stereotypes to justify their selfish behavior.

We find participants on average do not distort their *misogyny belief*. In both comparisons the observed average treatment effects are small and insignificant. Using *Control Self-interest* as a reference group, we observe average treatment effects of 0.239. Using *Control Context* as the reference group, the treatment effects reach a slightly larger value of 1.129. To study if the treatment effects vary, we take a closer look at the distribution and variance of the beliefs and conduct quantile regressions. We can not reject the hypothesis that the *misogyny belief* in the *Main Treatment* has the same variance and distribution as in the two controls. In the quantile regressions we observe that the coefficients of the treatment dummies develop in a

4. Using answers to the IAB-BAMF-SOEP Survey of Refugees in Germany we are able to quantify this belief. Thus, distorting their belief is personally costly for the participants. We used the quadratic scoring rule to incentivize this question. Details on how the question has been introduced can be found in Section 2.

5. The pro-immigration organization was PRO ASYL.

6. The two pro-immigration organizations were PRO ASYL and BumF (Bundesfachverband unbegleitete minderjährige Flüchtlinge)

7. In *Main Treatment* the donation decision directly affected refugees and the belief was about the refugees' image of women, in *Control Context* this connection between donation decision and belief no longer exists.

similar fashion for both controls. For a long time - up to the 75th percentile - the treatment effects stay with one exception insignificant and are also relatively small in magnitude. At the 85th percentile we begin to observe significant effects: the *misogyny belief* is 4.752 and 4.421 points higher in *Main Treatment* than in *Control Self-interest* and *Control Context* respectively. Together with the average treatment effects the observed heterogeneity in the response to the treatment suggests that only a small subset of participants with the most extreme beliefs about refugees might be willing to distort their *misogyny belief* to rationalize their behavior.

To fully comprehend the reason underlying the lack of motivated belief updating in the rest of the population, we run several additional tests. All of the above analyses rely on a crucial assumption that participants, in all three experimental conditions, answer the two questions on the decision page simultaneously - meaning they have one question in mind when answering the other and vice versa.[8] If this is not the case and participants look at the two decisions in isolation, we should not expect participants to distort their *misogyny belief* in a motivated manner. To test if participants answered the questions in the desired way, we exploit the randomized order of the two questions on the decision page. If participants made the two decisions simultaneously the order of the two questions on the decision page should not have any effect on the answers. Yet, we observe large and significant order effects in *Main Treatment*.[9] When the *misogyny belief* question was displayed after they were able to take money away from the donation to the pro-immigrant organization, the *misogyny belief* in *Main Treatment* was higher by 10.6%. This directly contradicts our assumption that participants answered the two questions on the decision page simultaneously and explains the missing average treatment effects. In the subsequent step, we present evidence that the participants answered the two questions sequentially, i.e. they answered one question after the other.[10] A sequential answering mechanism opens the door for a different kind of belief distortion: when participants get the donation decision displayed atop they might retroactively adjust their *misogyny belief* to justify their past behavior. To present support for this, we run the average treatment regression for the two order subsamples separately. When the *misogyny belief* question was displayed below the donation decision, we observe positive and for *Control Context* significant treatment effects. Using *Control Context* the *misogyny belief* is by 5.539 points higher in the *Main Treatment* and using *Control Self-interest* the belief is by 4.711 points higher. When the *misogyny belief* question was asked before the donation decision

8. To ensure this, participants had to spend at least 60 seconds on the decision page and, additionally, the order of the two questions on the decision page was randomized.

9. We observe no order effects in our two controls.

10. Besides behavioral measures (time spent and clicks on the decision page) we observe that the correlation between donation and *misogyny belief* in *Main Treatment* is larger when the belief question was displayed below the donation.

the coefficients are negative, smaller in magnitude and always insignificant. This evidence suggests that participants that answered the *misogyny belief* after the donation indeed exhibit some kind of ex-post rationalization of their selfish behavior.

Our research relates to the broader literature on motivated reasoning, with a special focus on research that studies how individuals are motivated to justify selfish or immoral behavior by distorting their beliefs. Further, it adds to the literature on the nature and function of stereotypes.

Motivated reasoning is an active field of research and has been for years (see, e.g. Kunda (1990) and Epley and Gilovich (2016)). Of particular interest to this paper is the literature on motivated reasoning in the domain of social preferences. Individuals like to think of themselves as charitable and selfless.[11] However, quite frequently individuals do not behave as generously or morally as they would like to perceive themselves to be. Our findings relate most closely to papers that study how individuals rationalize such selfish actions by distorting related beliefs.[12] In this context evidence indicates that individuals distort beliefs about how other people behave (see, e.g. Di Tella et al. (2015) and Falk, Neuber, and Szech (2020)), risk preferences (see, e.g. Exley (2016)), preferences over fairness (see, e.g. Konow (2000) and Dana, Weber, and Kuang (2007)), ambiguity preferences (see, e.g. Haisley and Weber (2010)), beliefs about the charity performance score (see, e.g. (Gneezy, Keenan, and Gneezy, 2014; Exley, 2020)) and beliefs about a product's quality (see, e.g. Chen and Gesche (2017) and Gneezy, Saccardo, Serra-Garcia, and Veldhuizen (2020)). In contrast to these studies, our paper tests if individuals distort stereotypical beliefs about a social group to rationalize selfish behavior. Holding and expressing stereotypical beliefs about marginalized groups is in itself already undesirable and could also have dire consequences for the group. Thus, we look at an extreme case of motivated belief distortion and shine a light on the question of how far individuals are willing to go to justify their selfish behavior.

Our paper also relates and contributes to the literature on stereotypes. According to Bordalo, Coffman, Gennaioli, and Shleifer (2016) there are three broad approaches

11. Other contexts researchers have also focused on are motivated reasoning in regards to: ability and beauty (see, e.g. Eil and Rao (2011), Mobius, Niederle, Niehaus, and Rosenblat (2011), Grossman and Van der Weele (2017), Coutts (2019), Schwardmann and Van der Weele (2019), and Zimmermann (2020)) or politics (Thaler, 2019). Bénabou and Tirole (2016) present an overview over this literature. For theoretical work on motivated reasoning see Bénabou and Tirole (2002) and Köszegi (2006). There is also a considerable literature on the mechanics of motivated beliefs, i.e. how individuals in light of contrasting actions or feedback maintain the desired beliefs. Zimmermann (2020) studies the role of memory, while other researchers have pointed toward motivated decision errors (Exley and Kessler, 2019), self-serving attributions to an unknown state of the world (Heuser and Stötzer (2020)), information avoidance (see, e.g. Dana, Weber, and Kuang (2007) and Golman, Hagmann, and Loewenstein (2017)) or conservatism and asymmetry (see, e.g. (Eil and Rao, 2011; Mobius et al., 2011)).

12. For an overview of this literature see Gino, Norton, and Weber (2016)

to stereotypes in social science. The first approach is called statistical discrimination and is based on the work of Phelps (1972) and Arrow et al. (1973). The theory of statistical discrimination argues that stereotypes are based on rational expectations. This means beliefs about a certain member of a group are formed by the aggregated distribution of group traits. However, statistical discrimination misses one point: stereotypes are rarely accurate. The second approach, which has a long tradition in social psychology (see, e.g. Schneider, Hastorf, and Ellsworth (1979) and Schneider (2005)), is called the "social cognition approach". According to this approach, stereotypes are cognitive schemes. Individuals save cognitive resources, by defining others over representative attributes of the group they belong to. Attributes of a group are representative if they stand out compared to other groups, e.g. people from Florida are old or Dutch people are tall. Thus, stereotypes have a kernel of truth to it, but at the same time can lead to misjudgments (see, e.g. Judd and Park (1993) and Bordalo et al. (2016)). A third approach views stereotypes as a pejorative generalization of group traits. According to this approach, stereotypes always serve a purpose for people that are applying them (see, e.g. Glaeser (2005)). Our paper most closely relates to the third approach. We hypothesized that participants distort stereotypical beliefs about refugees in order to earn money at refugees' expense. In addition to the literature investigating the reason that people form stereotypes, Kundra and Sinclair (1999) present evidence on how motivated reasoning and stereotypes are intertwined using predominantly observational studies that show how motives lead to both activation and application of stereotypes. Our findings are, to the best of our knowledge, the first to present causal evidence of whether individuals activate stereotypical beliefs about a group to rationalize their own selfish behavior.
The remainder of the paper is structured as follows. We first introduce the experimental design and procedure. Section 2.3 presents our empirical strategy and results, which are divided into two subsections: first, an investigation of average treatment effects and subsequently one on quantile regressions. Section 2.4 presents a discussion of the observed null results from the previous section. Section 2.5 concludes.

2.2 Experimental Design

2.2.1 Logistics

We implemented an online survey experiment using a nationally representative sample of 902 adult citizens in Germany. The sample was provided by Pureprofile, a large market research company, and the experiment was computerized using the Qualtrics online survey tool. The sample is representative of the adult German population with respect to age, gender, the region of residence, income and graduate

population.[13] On average the participants spent approximately 7 minutes and 30 seconds answering the survey.
The survey experiment consists of roughly 3 components: (i) An attention check to ensure high quality responses, (ii) a screen that elicited demographics and routed participants to one of the experimental conditions, depending on whether they matched the desired sample characteristics, and (iii) a decision page to measure participants' *misogyny belief* and willingness to take money away from a donation.
Two questions were presented on the decision page. To encourage participants to answer the questions carefully and simultaneously - meaning having one question in mind while answering the other and vice versa - we additionally implement a requirement that participants spent at least 60 seconds on the decision page, as well as a randomization of the order in which the the questions were presented.
To ensure high quality responses, all participants had to pass an attention check wherein they were asked to give prespecified answers to a trivial question.[14] If a subject failed to pass the attention check, they were redirected to the company website and were not allowed to participate in the experiment. Participants who passed the attention check went on to complete a questionnaire that elicited demographic information, which was in part used to ensure the representativeness of our sample. If a representativeness quota such as the amount of people between 18 - 25 years old, was already fulfilled, a new subject in this category was redirected to the survey companies' website and was not allowed to participate in the experiment.[15]
Besides the quota relevant variables (age, gender, income, region and education), we elicited data on participants' self-placement on a left-right political spectrum, migration background, type of residential area and religion. Table 2.A.1 shows that the different treatment groups are balanced in the observable characteristics.
In the next subsection we describe the different treatment conditions and the elicited variables.

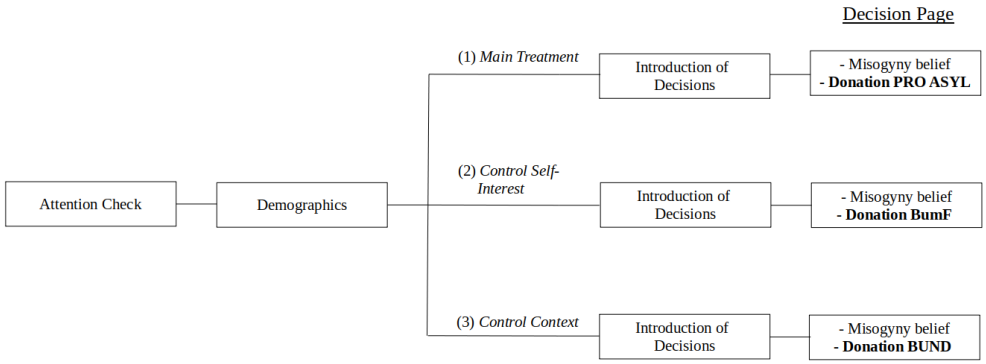
2.2.2 Survey Experiment

An environment to study whether participants strategically use stereotypes requires (a) a context in which individuals might be motivated to distort their beliefs about a certain group and (b) control conditions in which the motives to distort beliefs are cut off. Our design accommodates both features. Figure 2.1 illustrates the experimental procedure.

13. Both age and income were divided into intervals. Region of residence in our context means whether participants come from East or West Germany. Graduate population is the percentage of individuals that hold a degree from an institution of higher education.

14. The attention check page can be seen in Appendix 2.B.1 and asks participants to give prespecified answers to the following question: Are you interested in Game of Thrones (a famous TV show)? Only 32 % gave the right, meaning the prespecified, answers to the question.

15. We did not have to pay for these "surplus" respondents.



The figure shows our experimental design. After passing an attention check and answering a short survey, participants were randomly assigned to one of three treatment conditions. All experimental conditions have an identical structure where the only difference was the donation decision.

Figure 2.1. Experimental Design

After successfully passing the attention and demographics screens, each participant was randomly allocated to one of three different experimental conditions: *Main Treatment*, *Control Self-interest* or *Control Context*. The different conditions all followed the same structure: participants were given general information about the study and the payment structure, and were introduced to the two decisions they subsequently had to make on the decision page.[16] The decision page itself consisted of two questions. While one of the two decisions (*misogyny belief* question) stayed the same, the second decision (donation decision) varied across the three experimental conditions. We now turn to the two decisions:

Main Outcome: *misogyny belief*

As outlined in the introduction, we seek to study whether participants distort their beliefs about refugees that arrived in Germany during the 'European migrant crisis' in a motivated manner. To do so, we elicit participants' belief about the refugees' attitude towards equal rights for women. Participants were asked:

"*Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy?*"

We will call a participant's answer to this question: *misogyny belief*.
We can quantify this belief using the answers to the IAB-BAMF-SOEP Survey of

16. The payment was structured as follows. In addition to the fixed payment, one out of 10 participants was randomly chosen for additional payments. If a participant was chosen, one of her two decisions was implemented. We informed participants at the end of the survey whether they were chosen and, if so, how much they and others additionally earned.

Refugees in Germany, which was representative for refugees arriving in Germany between 2013 and 2016. Refugees were asked to state, on a scale from 0 (in no case) to 10 (absolutely), how much they think equal rights for women should be part of every democracy. We used the refugees' answers to this question to quantify our *misogyny belief* question.[17] Before the decision page, participants received information about the origin of the data, the question they had to answer and how they could earn money. In all three conditions, both the introduction to as well as the actual question regarding the *misogyny belief* were identical. Incentive compatibility was ensured by using the quadratic scoring rule. Besides showing participants the exact payout formula, they were told that the closer their guess is to the true value, the more money they could earn.[18]
As we will explain in the next paragraph, the difference between the treatment and control conditions lies in the extent to which the donation decision was connected to the *misogyny belief*.

Experimental Conditions: different versions of donation decision
(1) Main Treatment In the *Main Treatment* condition, aside from stating their belief about refugees' attitudes towards equal rights for women (*belief misogyny*), participants had the opportunity to take money away from a donation to one of the largest pro-immigration advocacy groups, PRO ASYL. *Main Treatment* fulfills all the necessary conditions needed to study whether participants are motivated to distort their beliefs or not. Participants can enrich themselves at the expense of refugees and simultaneously have the chance to distort their belief about the refugees' beliefs about women rights.
In line with the belief decision, participants received all relevant information about the donation decision before the decision page. More specifically participants were told that 50 EUR would be donated to PRO ASYL and that they have the opportunity to take money away from the donation and keep it for themselves. For every 1 EUR

17. Only 1.6 % of the refugees in the IAB-BAMF-SOEP Survey stated that in *no case* women should have equal rights. For our *misogyny belief* question we used the exact wording from the scale and asked participants whether women 'in no case' should have equal rights. 96.4 % of the refugees stated a number 5 or higher on the 10 point scale and 78.6% of the refugees said that women should absolutely (10) have equal rights. It is important to highlight that in the original question, refugees were asked to state their answer on a 10 point Likert Scale as opposed to a binary answer. Thus, the discrepancy between the 'true answer' (1.6 %) and the beliefs of the participants - only 8 % of the participants report a belief equal or lower than 10 out of 100 - might, to some extent, be to the modification of the question. But as we are interested in how the different conditions alter the beliefs, we see the discrepancy to be unproblematic.

18. The exact formula is

$$P(\text{misogyny belief}) \;=\; 2 - 2 * 0.02(\frac{guess}{100} - \text{true value})^2$$

where the true value is 0.016 - meaning 1.6 % of the people in the IAB-BAMF-SOEP Survey of Refugees in Germany said that women should in no case have equal rights.

they take away, participants could receive 50 cents. Further, participants received information about PRO ASYL and their general cause.[19]

(2) Control Self-interest In *Control Self-interest* participants were asked to shift money between two pro-immigration advocacies. Thus, participants were neither able to destroy money nor take money for themselves. By ruling out the possibility to enrich themselves, participants should have no interest in distorting their answer to the *misogyny belief* in a motivated manner.
Prior to the decision page, participants were told that we would donate 50 EUR to PRO ASYL but that they have the chance to shift some money from the donation to BumF (Bundesfachverband unbegleitete minderjährige Flüchtlinge). Additionally, participants received information about the two organizations.[20] For every 1 EUR participants took away from the donation to PRO ASYL, 1 EUR was given to BumF.

(3) Control Context In the second control condition (*Control Context*), participants decided whether they wanted to take money away from a donation to a German environmental organization called BUND (Bund für Umwelt und Naturschutz Deutschland) in order to enrich themselves. In contrast to *Main Treatment* there no longer exists a connection between the stated belief (*misogyny belief*) and the donation decision. By cutting off the relationship between the two questions, participants are no longer able to rationalize their donation behavior by distorting their misogyny beliefs.
In line with the procedure for the other experimental conditions, participants received information about the organization, were told that we would donate 50 EUR to BUND and that they have the chance to take money away from this donation.[21] Again, for every 1 EUR participants take away, they receive 50 cents.

19. We showed participants the following quote from the PRO ASYL website: "PRO ASYL advocates for the rights of refugees in Germany and in Europe. We help them to apply for asylum. We investigate human rights violations. And we campaign for an open society in which refugees receive protection." see Pro Asyl (2020).

20. As in *Main Treatment* we showed participants the following quote from the PRO ASYL website: "PRO ASYL advocates for the rights of refugees in Germany and in Europe. We help them to apply for asylum. We investigate human rights violations. And we campaign for an open society in which refugees receive protection." see Pro Asyl (2020). For the BumF we showed participants the following quote "Since 1998 the Association for Unaccompanied Refugee Minors (Bundesfachverband unbegleitete minderjährige Flüchtlinge: BumF) advocates for the rights of displaced children, adolescents and young adults in Germany. [..] It is our aim, that young refugees grow up without fear, marginalization or discrimination and enjoy the same rights as any other young person." see Bundesfachverband unbegleitete minderjährige Flüchtlinge (2020)

21. Analogous to the instructions in the other conditions, the participants received information about the organization from their website "BUND advocates for the protection of our nature and environment - so that the earth is habitable for all that live on it". This is our own translation - for the German text see Bund für Umwelt und Naturschutz Deutschland (2020).

2.3 Results

Before turning to our results, we outline our empirical strategy for identifying average treatment effects. Comparing the *misogyny belief* between the *Main Treatment* and the two controls enables us to make causal inferences about the motivated use of stereotypes. As stated before, participants in *Main Treatment* could rationalize taking away and destroying money from a donation to refugees by distorting their *misogyny belief*. The two control conditions aim to rule out such motivated reasoning. However, they do so in different ways. In *Control Self-interest* participants can't receive money from the donation decision and therefore can't act in a self-interested manner. The missing connection between beliefs about refugees and donations to BUND in *Control Context* removes the possibility for participants to rationalize their behavior. As the two control conditions differ in their way of ruling out motivated belief distortion, we compare the treatment effects for the two control conditions separately. We run the following regressions:

$$Y_i = \alpha + \beta_{self-interest} * treat_{self-interest,i} + \gamma * Controls_i + \epsilon_i \quad (2.1)$$

$$Y_i = \alpha + \beta_{context} * treat_{context,i} + \gamma * Controls_i + \epsilon_i \quad (2.2)$$

where Y_i denotes our dependent variable *misogyny belief*. The variable of interest, $treat_{self-interest}$ ($treat_{context}$), is a dummy variable indicating whether participants were in *Main Treatment* or in *Control Self-interest* (*Control Context*).
For each control group, we run two OLS-regressions - one without and one with control variables. The $Controls_i$ added to the OLS-regression are dummies for age groups, gender, region of residence, self-placement on a left-right political spectrum, income, university degree, migration background, urban area and Christian denomination.[22]
As highlighted before, we argue that participants in the *Main Treatment* want to pocket some of the donation money for themselves. Taking money from a disadvantaged group would however be in stark contrast with a positive self-image, thereby motivating participants in the *Main Treatment* to distort the *misogyny belief* upwards. Running the described regressions enables us to test this hypothesis:

H1 Participants distort their *misogyny belief* in a motivated manner when given the opportunity to enrich themselves at the expense of refugees.

H1.1 For (2.1) this means $\beta_{self-interest} > 0$

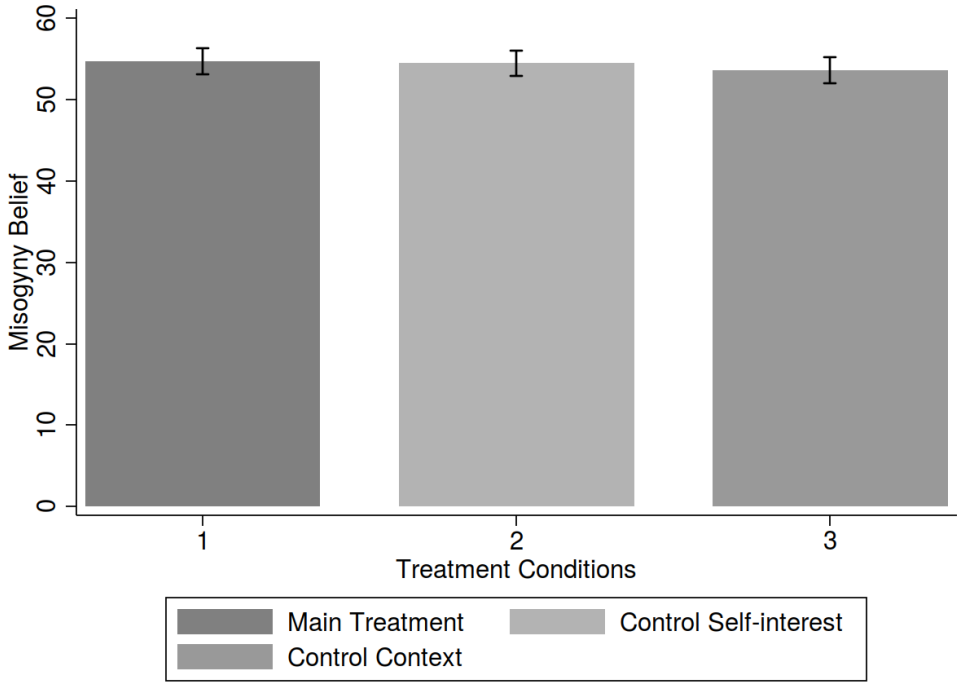
H1.2 For (2.2) this means $\beta_{context} > 0$

The upcoming results section is divided into two parts: First, we look at average treatment effects. Second, using quantile regressions, we explore the heterogeneous effects of the treatment.

22. The age groups are 18-25, 25-45, 45-65 and > 65. Region of residence is a dummy indicating whether the participant lives in East Germany.

2.3.1 Average Treatment Effect

As outlined in the previous paragraph the *misogyny belief* should be significantly higher in *Main Treatment* than in the two controls. However, as we can see in Figure 2.1, there exists virtually no difference in the *misogyny beliefs* between the treatment conditions. Our results in Table 2.A.2 confirm this finding. In column (1)



The figure plots the mean *misogyny belief* for each experimental condition. The left bar plots the average *misogyny belief* in *Main treatment*. The other two bars plot the average *misogyny belief* in the two controls (middle bar: *Control Self-interest*, right bar: *Control Context*). The error bars indicate + / - 1 standard error.

Figure 2.1. Average *misogyny belief* in the experimental conditions

and (2) we observe the treatment effects using our first control *Control Self-interest*. In the regression without any controls (column (1)) the coefficient of the treatment dummy is 0.239. When we add the previously discussed control dummies, the coefficient is 0.569. Both effects are statistically insignificant. The two remaining columns ((3) and (4)) show average treatment effects using our second control *Control Context*. Although the magnitude of the treatment coefficients is larger, 1.129 and 1.608, respectively, the treatment on average had no significant impact on the *misogyny belief* of the participants.

In summary on average participants do not distort their beliefs about refugees in a motivated manner, which is to say that on average participants not use stereotypes

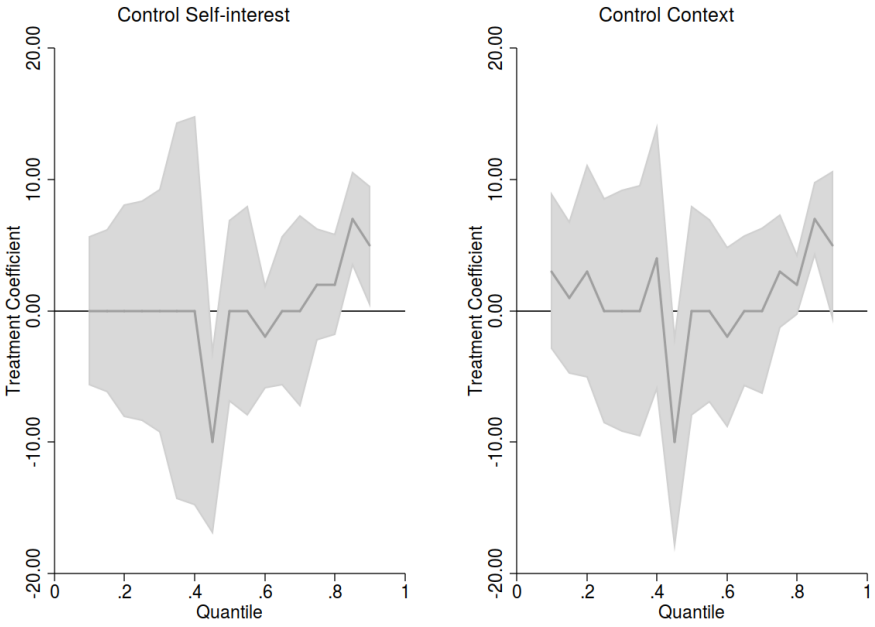
as a tool to rationalize selfish behavior.

2.3.2 Treatment Effect Variation

The previous section revealed that on average participants did not distort their *misogyny belief* in *Main Treatment*. In this part, we look at variations of the treatment effect. We present evidence that the treatment pushed a subsample of participants with already high stereotypes to state even higher beliefs.
Studying the distributions and variance of the *misogyny belief* in the different groups provides us with first indicators of heterogeneous treatment effects. Figure 2.A.1 plots the distributions plus respective kernel density and Figure 2.A.2 the cumulative distributions of the *misogyny belief* for the three different experimental conditions. Testing for equality between the *Main Treatment* and respective control, we cannot reject the hypothesis that they follow the same distribution. Comparing the variance of the *Main Treatment* and respective control paints an identical picture. The variance in beliefs does not differ significantly. But when we focus on extreme values in Figure 2.A.1 and 2.A.2 we observe that there seems to be far more mass in *Main Treatment* than in the two controls. We run several quantile regressions to test this argument. Figure 2.2 and Tables 2.A.6 and 2.A.7 report our findings. While Figure 2.2 plots the estimate for the two treatment dummies and the respective 95% confidence interval without control dummies, the two tables report our findings for a selection of percentiles using controls.[23] Both the figure and the two tables reveal a similar pattern: Beginning around the third quartile, the coefficients become larger in magnitude and eventually significant (at least at 5%).[24] Using *Control Self-interest* (Table 2.A.6), we find that at the 85th percentile the coefficient of the treatment dummy is 4.752 (significant at 5%). Similarly, for our second control (see Table 2.A.7) the 85th percentile in *Main Treatment* is higher by 4.421 points (significant at 1%). Taken together, this indicates that only those participants with already high stereotypes distorted their *misogyny belief* due to the treatment. In the final step, we want to find out who holds such high *misogyny belief*. To see what observables are especially predictive of having a high stereotype, we run a simple probit regression with a high stereotype dummy variable as the outcome

23. Namely the 15th percentile, the first, second and third quartile, and the 85th percentile. Additionally,we report the average treatment effects, as a benchmark in column (1).

24. The two plots in Figure 2.2 look strikingly similar. Besides significant and positive treatment dummies for high percentiles, we observe a sharp fall in the coefficient around the 45th percentile. Responsible for the observed effects is the distribution of answers for participants that stated a *misogyny belief* between 45 and 60. While participants in *Main Treatment* are bundled around 50 in the two Controls they are evenly spread, leading to the observed spike and significant negative treatment effects.



The figure shows the quantile treatment effects on the *misogyny belief*. It depicts the estimated quantile treatment coefficients of our specification without controls and the 95 % confidence interval from the 10th to the 90th percentile. For the left picture we used *Control Self-interest* and for the right picture we used *Control Self-interest* as the reference group.

Figure 2.2. Quantile Regression

and all our observables as independent variables.[25] The results of the regression are reported in Table 2.A.5. We find that the following characteristics have a significant impact on the probability of having a high stereotype: age, self-classification on a left-right spectrum and whether participants have a university degree. While being older and classifying oneself as more to the right on the political spectrum raise the likelihood of having a *misogyny belief* above the third quartile, having a university degree has a significant negative effect on it.

Our results show that participants do not on average distort their *misogyny belief* in a motivated manner. The quantile regressions in Subsection 2.3.2 revealed that only a small part of participants, i.e. those with already high stereotypes, seem to rationalize their donation behavior by altering their *misogyny belief*.

25. The stereotype dummy equals one if participants state a *misogyny belief* above the 75th quantile of the whole sample.

2.4 Discussion

The observed results from Section 2.3 support two possible conclusions. Either individuals are not willing to 'use' stereotypes to rationalize selfish behavior, or underlying assumptions of our experimental design were violated. In the following we will focus on the second conclusion. We explicitly test the two central assumptions of our design: (i) There has to be a thematic connection between the donation decision and *misogyny belief* in the *Main Treatment* and (ii) Participants have to answer the two questions on the decision page simultaneously.

Assuming participants do not see a thematic connection between the donation to *PRO ASYL* and the *misogyny belief*, distorting the belief is no solution to rationalizing the donation behavior. To show that this condition is fulfilled, we run two regressions using only the participants in *Main Treatment* - one without and one with controls. A strong correlation between the *misogyny belief* and the behavior in the donation decision would indicate that there is a thematic connection between the two questions.[26] Table 2.A.3 reports the findings. We see that lower levels of donations are significantly (at 1%) correlated with higher *misogyny beliefs*. In our specification without controls, a 10 points higher *misogyny belief* leads on average to a 2.11 EUR lower donation. Importantly, this connection does not exist for participants in *Control Context*.[27] Thus, in showing that within the *Main Treatment* - and only there - lower donations are significantly correlated with higher stereotypes about refugees, we demonstrate that assumption (i) of our design was indeed fulfilled.

In a subsequent step, we check if assumption (ii) holds. If participants answered the questions in isolation instead of simultaneously, i.e. a participant formed a belief without thinking about the actions in the donation decision, we shouldn't expect motivated reasoning to occur. To test if this assumption holds, we leverage the randomized order of questions on the decision page. We observe that the order of the questions affects the decisions and therefore conclude that participants did not answer the two questions simultaneously. This is a violation of a central experimental design assumption and might explain the absence of treatment effects. In what follows we present how the order of questions on the decision page affected our results.

26. While a correlation between the donation decision and the *misogyny belief* is desired in *Main Treatment*, it is not in *Control Context*. To show that there is no connection, we run the same two regressions using only the participants in the *Control Context*. As the donation decision differed greatly in *Control Self-interest* and participants were not able to take away and waste money, we do not report the findings here.

27. Table 2.A.4 shows that higher *misogyny beliefs* are generally associated with lower donations. However, the correlation is not significant and not nearly as strong as for participants in *Main Treatment*.

2.4.1 Simultaneous decision making

A central assumption of our design was that participants answered the two questions on the decision page simultaneously - meaning having the one question in mind while answering the other and vice versa. If participants instead answered the first question on the decision page without regard to the second one, motivated reasoning should not occur. To test this assumption, we utilize the randomized order of the two questions on the decision page. As previously outlined, participants either saw the *misogyny belief* question above the respective donation decision or vice versa.[28] When participants make their decisions simultaneously, the order on the decision page should have no impact on the answers to the *misogyny belief* question.[29] In the following paragraphs, we first show how the order of the two questions affected the *misogyny belief* in the treatment and control groups in more detail. After having shown that the central assumption failed, we explore how the order affected the regression results from Section 2.3. In the final step, we present evidence that participants answered the questions on the decision page sequentially, i.e. one after the other, and show the consequences of such an answering mechanism.

Effect of display order on *misogyny belief*. To test whether participants answered the two questions on the decision page simultaneously, we create a dummy variable called *belief_first* that equals one if the *misogyny belief* question was shown above the other question and zero if the donation decision was shown first. For all three experimental conditions (*Main Treatment* plus the two control groups) we run separate OLS-regressions with *misogyny belief* as our outcome variable and *belief_first* as the variable of interest. Table 2.1 reports our findings. The first two columns show our results for *Main Treatment*, columns (3) and (4) for *Control Self-interest* and the last two columns for *Control Context*. The results in Table 2.1 reveal that, depending on the order of the two decisions, the participant stated significantly different beliefs in our *Main Treatment*. When the computer displayed the *misogyny belief* question after the donation decision ($belief\_first = 0$), participants stated a significantly higher belief. The *misogyny beliefs* were higher by 6.148 and 7.503 points, which in the specification without controls reflects a 10.6% change in the stated belief. As expected, the order did not significantly affect the stated belief in both control groups (see columns (3) to (6)). This is compelling evidence that in our control groups no connection between the behavior in the donation decision and the *misogyny belief* existed and therefore the motive to distort the belief was indeed absent. Nevertheless, it is important to point out

28. Picture 2.B.2 is a screenshot of the decision page in which the *misogyny belief* question was randomly displayed above the donation decision. Picture 2.B.3 shows the other case in which the donation decision was shown first.

29. Or, if participants always answer one question without regard to the other and vice versa, the order should also do not affect the answers.

that the coefficients of *belief_first* go in the opposite direction. In both controls, the *misogyny belief* is higher when the respective donation decision was shown first.

Table 2.1. Order Effects

	Main Treatment		Control Self-interest		Control Context	
	(1)	(2)	(3)	(4)	(5)	(6)
belief_first	-6.148*	-7.503**	2.697	2.609	3.083	4.607
	(3.229)	(3.397)	(3.095)	(3.226)	(3.202)	(3.353)
Constant	57.73***	27.05	53.02***	32.05**	52.19***	73.35***
	(2.196)	(32.08)	(2.201)	(14.96)	(2.163)	(16.97)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.0120	0.241	0.00252	0.188	0.00309	0.190
R^2_a	0.00867	0.110	-0.000824	0.0444	-0.000257	0.0462
N	302	302	300	300	300	300

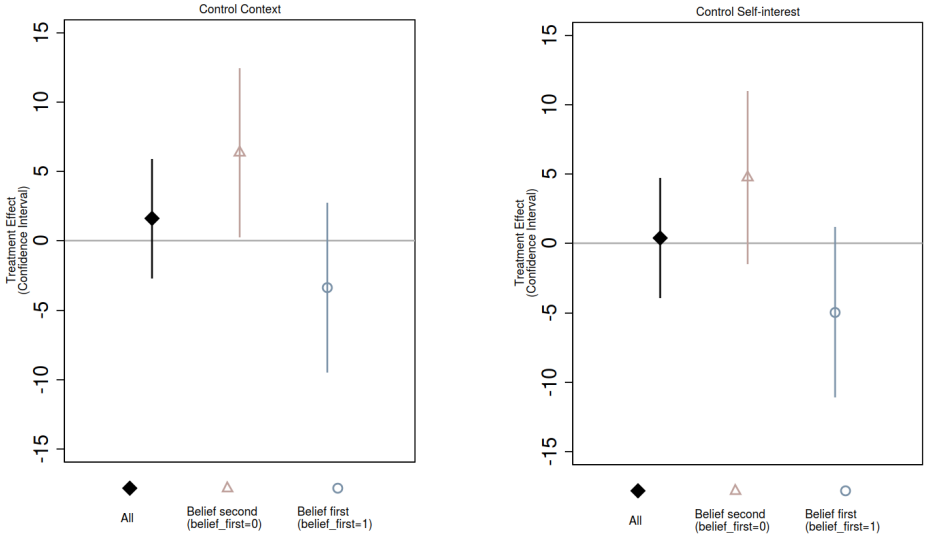
Notes: This table reports the effects the display order of the two questions on the decision page had on the *misogyny belief*. We look at all three experimental conditions separately. To study the effects of the order, we construct a dummy variable *belief_first*, which equals 1 if the *misogyny belief* question was displayed above the respective donation decision and equals 0 if the respective donation decision was displayed above the *misogyny belief* question. Columns (1) and (2) report our findings for *Main Treatment*. Columns (3) and (4) report our findings for *Control Self-interest*. Columns (5) and (6) report our findings for *Control Context*.
Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right political spectrum, migration background, Christian denomination and whether participant live in a urban area or not (urban). * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

The results from Table 2.1 show that participants in *Main Treatment* experienced the experiment differently depending on the order of the questions on the decision page. This directly contradicts our central design assumption (ii). We must therefore conclude that participants in *Main Treatment* did not answer the questions on the decision page simultaneously.

Effect of display order on regression results. Having established that a central design assumption did not hold, we now want to see how this affected the results from Section 2.3. Figure 2.1 shows how the coefficient of the respective treatment dummy varies depending on the order of the two questions.[30] In both pictures

30. In addition to the two pictures, we report our findings in two tables. Table 2.A.8 and Table 2.A.9 document how the treatment effects from Section 2.3.2 differ depending on the order of the two questions. Both tables follow the same structure: In column (1) and (2) we introduce an interaction term, treatment dummy times *belief_first* dummy, in columns (3) and (4) we focus on the subsample

we plot the coefficients and the 95% confidence interval for the three samples: the square and line report our findings for the whole sample (treatment and respective control group), the triangle and line report the finding for participants that saw the donation decision first (*belief_first* = 0) and the point reports our findings for the subsample of participants that saw the *misogyny belief* question first (*belief_first* = 1).



The figure shows the treatment effects on the *misogyny belief*. It depicts the estimated treatment coefficients of our specification with controls and the 95 % confidence interval from the 10th to the 90th percentile. The left-hand side picture reports our findings for *Control Context* and the other picture for *Control Self-interest* as the respective reference group to *Main Treatment*. In both pictures, the square reports the treatment coefficient and the line the 95 % confidence interval for the whole sample. Analogously, the triangle and the attached line report the results for the subsample for which the donation decision displayed above the *misogyny belief* question (*belief_first* = 0) while the point and the corresponding line report the results for the subsample for which the *misogyny belief* question was displayed first (*belief_first* = 1).

Figure 2.1. Effect on *misogyny belief* for order subsamples

In the first picture, *Control Context* is the comparison group. As previously reported, we observe no average treatment effects. Focusing on the two subsamples, we observe that the coefficients are in opposing directions. Participants who saw the *misogyny belief* question after the donation decision exhibit significant treatment ef-

which observed the donation decision first (*belief_first* = 0) and in the last two columns ((5) and (6)) we report our findings for the subsample for which the *misogyny belief* question was displayed first. The interaction term in column (1) and (2) is significant in both tables, showing that, depending on the order, participants reacted differently to our treatment. Table 2.A.8 reports our findings using *Control Self-interest* as the reference group.

fects: The *misogyny belief* in *Main Treatment* is on average 6.359 points larger when the donation decision was displayed above the belief question. This effect is significant at 5%. Although not significant, participants in *Main Treatment* who saw the belief question first state −3.365 lower *misogyny belief*. The second picture uses *Control Self-interest* as the reference group and tells an identical story. The coefficient of the treatment dummy ($treat_{self-interest}$) is positive for the one subsample and negative for the other. The coefficient of $treat_{self-interest}$ is 4.607 when the *misogyny belief* question was asked after the donation and −3.332 when the order was the other way around. However, the observed effects are not significant.
At this stage, we are certainly cautious in interpreting these findings in terms of our hypothesis. Nevertheless, the results from the regressions are further proof that the order of the two questions had a substantial effect on the participants. In the next paragraph, we present a mechanism that will help to apprehend above findings.

Alternative mechanism: Sequential decision making. In the final step, we want to investigate what the mechanism behind the observed order effects is. After corroborating that the order of the two questions has a significant effect on the answers in *Main Treatment* and, as a result, affect the coefficients of our regression, we now seek to deepen our understanding of the reason behind these effects. We argue that participants answered the two questions sequentially. This means they answered the first question without thinking about the second. Thus, instead of interaction effects between the two questions, we should only observe unidirectional effects from the behavior in the first to the second question.
What distinguishes sequential decision making from other explanations for the observed order effects is that participants, in answering one question after the other, essentially behaved similarly in both settings.[31] We proceed to show that neither the optics of the decision page nor the click- and time-spent-behavior differs greatly between the two orders. We interpret this as evidence in favor of sequential decision making. Figures 2.B.2 and 2.B.3 in Appendix B show the decision page for the two different orders. Importantly, in both pictures, the two questions can be seen at the monitor at the same time.[32] Thus, the optics of the decision page did not differ greatly depending on the order. Further, to see how the order affected the time spent and the number of clicks on the decision page in *Main Treatment*, we run several regressions with *belief_first* as the variable of interest. Table 2.A.10

31. Other potential explanations would assume that participants acted differently depending on the order, e.g. when the *misogyny belief* question was displayed second, participants answered simultaneously, but when the donation question was displayed second, participants looked at the questions in isolation.

32. This should be the case for most standard displays. However, we can not rule out that participants used smaller devises, like a smart phone or a tablet. Thus, we clicked through the survey with different devices and found that for for a 10 inch display, both questions could be seen, whereas with a smartphone with a 5 inch display, only one question could be seen at a time.

reports how the order affected participants' behavior on the decision page. As none of the regressions indicate that the order (*belief_first*) has a significant effect, we conclude that the order did not systematically alter the participants' behavior - outside the answers to the two decisions - on the decision page. Additional support for sequential decision making is presented in Table 2.A.11, which reports the correlation between the behavior in the donation decision and the *misogyny belief* for the two order subsamples in the *Main Treatment*. In support of the proposed mechanism, higher *misogyny beliefs* are more strongly correlated with lower donations when $belief\_first = 0$.
We argue that the presented suggestive evidence points towards sequential decision making, i.e. participants in all settings answered the question that got displayed on top first and only afterwards thought about their answer to the second question. As stated before, a sequential decision process would only have spillovers from the first to the second decision, i.e. if $belief\_first = 0$, participants answered the *misogyny belief* question keeping their previous behavior in the donation decision in mind. This could explain the higher *misogyny beliefs* in *Main Treatment* when the belief question was displayed after the donation and also the significant treatment effects (using *Control Context*) when $belief\_first = 0$. Participants who take money away from the donation are aware that such behavior is questionable and, by stating a higher *misogyny belief*, try to ex-post rationalize their behavior. Such a rationalization strategy is not possible in the setting where participants first committed to a belief and then acted upon this belief.[33]

We showed that a central assumption of our design was violated. Participants do not answer the two questions on the decision page simultaneously. Without simultaneous decision making, we should not expect motivated reasoning. As a consequence, the observed null results in Section 2.3 can not be seen as conclusive evidence that people do not distort their stereotypes about refugees in a motivated manner. On the other hand, there appears to be evidence for sequential decision making and ex-post rationalizing of selfish behavior.

33. Johnson, Häubl, and Keinan (2007) argue that individuals construct values of endowment by asking themselves queries. They show that altering the order or emphasise of the queries could greatly affect the value placed on a good. A similar mechanism might be at play for our *misogyny belief*: When the *misogyny belief* was displayed first, participants seem not to incorporate their decision in the the donation task in their belief formation. In contrast, when participants first answered the donation decision, they asked themselves how their beliefs about refugees are aligned with their behavior towards the group. The order in which specific questions are asked seems to influence the things we account for and as a consequence, affects our construction of values or beliefs.

2.5 Conclusion

This paper uses a representative sample of 902 participants to study if individuals distort their beliefs about refugees in order to behave selfishly in situations where they could enrich themselves at the refugees' expense. To investigate this question, we conducted an online experiment in which each participant was randomly allocated to one of three treatment conditions. In all treatment conditions, participants estimate the percentage of refugees who said that women should in no case have equal rights in a democracy. We call a participant's answer *misogyny belief*. While the *misogyny belief* question stayed the same in all settings, a second decision varied: In *Main Treatment*, participants were able to take away money from a donation to a pro-immigration organization. In contrast, the connection between the *misogyny belief* question and donation decision was eliminated in both of our controls. In *Control Self-interest* the participants were not able to enrich themselves at all. In *Control Context* they were able to take away money from a donation to an environmental organization to give to themselves, thus, there is no clear connection between the belief about refugees and the selfish behavior in the donation decision. Comparing the *misogyny belief* between our *Main Treatment* and each control provides causal evidence on motivated stereotypes.

Our results show that on average participants did not distort their beliefs about the refugees' image of women (*misogyny belief*). Yet, our quantile analysis revealed that participants with high beliefs seemed to be pushed to state even more extreme beliefs in our *Main Treatment*. Section 2.4 highlights one downside of our design: in *Main Treatment* the order in which the two decisions were displayed on the decision page had a large and significant impact on the *misogyny belief*. Participants in *Main Treatment* that had to answer the donation decision first stated beliefs that on average were 6.148 and 7.261 percentage points larger. This is problematic as one central assumption in our setting is that participants have to make their decisions simultaneously. This means participants are supposed to answer the donation question while having the *misogyny belief* question in mind and vice versa. The discussed results show that this was not the case in our experiment. Instead, participants seem to answer the questions sequentially. Therefore, participants in *Main Treatment* who answered the donation first stated higher *misogyny beliefs* to ex-post rationalize their behavior. With sequential decision-making, this type of rationalization is not possible when the belief question was asked first.

In light of our findings, we have some suggestions for future work. Adapting the proposed design in ways that ensure simultaneous decision making could present conclusive evidence if participants distort stereotypical beliefs to rationalize selfish behavior. Further, we observe that participants that saw the *misogyny belief* after the donation decision used stereotypes to ex-post rationalize their previous behavior. Making it possible to distinguish between sequential and simultaneous decision-making might be a fruitful direction for future research.

One could argue that some individuals are motivated to distort their beliefs in the other direction, i.e. they are motivated to present themselves as more moral than they are by donating a lot and understating their resentments towards refugees. For us it is not possible to identify these individuals as participants could only destroy money in the experiment. Further, in Section 2.3.2 we saw no evidence that there is more mass for especially low *misogyny beliefs* in our *Main Treatment*. Nevertheless, adapting the proposed research design in a way that allows for motivated reasoning in both directions might provide new insights.

Another interesting direction for further work is the spread of stereotypes by participants that used them to rationalize their previous behavior. In their model Bénabou, Falk, and Tirole (2019) show that people who observe the immoral actions of a previous person can pass on the narrative, instigating social contagion. One could adapt the proposed design to study the spread and lasting consequences of distorted stereotypes.

Appendix 2.A Additional Tables & Figures

Table 2.A.1. Summary Statistics

	(1)		(2)		(3)		(4)	
	Whole Sample		Treatment		Control Self-interest		Control Context	
	mean	sd	mean	sd	mean	sd	mean	sd
Age	49.71	16.41	49.73	16.51	49.73	16.49	49.67	16.27
Male	0.50	0.50	0.52	0.50	0.49	0.50	0.48	0.50
Residents of East-Germany	0.20	0.40	0.19	0.39	0.18	0.39	0.22	0.41
left-right	3.77	1.33	3.85	1.35	3.67	1.20	3.80	1.41
Migration background	0.14	0.79	0.13	0.84	0.13	0.85	0.14	0.65
Income								
0 - 1300	0.19	0.39	0.19	0.39	0.21	0.41	0.17	0.37
1301 - 2600	0.33	0.47	0.31	0.46	0.36	0.48	0.33	0.47
2601 - 3600	0.26	0.44	0.27	0.44	0.23	0.42	0.28	0.45
3601 - 5000	0.18	0.38	0.19	0.39	0.16	0.36	0.19	0.39
$>$ 5001	0.05	0.21	0.05	0.22	0.04	0.20	0.04	0.20
Education								
Without prof. training	0.11	0.31	0.14	0.34	0.10	0.30	0.09	0.29
Vocational training	0.71	0.45	0.71	0.46	0.71	0.45	0.72	0.45
University degree	0.18	0.38	0.16	0.36	0.19	0.39	0.19	0.39
Denomination								
None	0.47	0.50	0.46	0.50	0.49	0.50	0.46	0.50
Christian	0.50	0.50	0.51	0.50	0.47	0.50	0.51	0.50
Islamic	0.01	0.09	0.01	0.08	0.01	0.10	0.01	0.10
Buddhist	0.00	0.05	0.00	0.00	0.00	0.06	0.00	0.06
Jewish	0.00	0.05	0.00	0.00	0.00	0.06	0.00	0.06
Hindu	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.00	0.07	0.00	0.06	0.01	0.08	0.00	0.06
No response	0.01	0.11	0.02	0.13	0.02	0.13	0.01	0.08
Christian Denomination								
Christian Denomination?	0.50	0.50	0.51	0.50	0.47	0.50	0.51	0.50
Area of Residence								
Farm	0.01	0.09	0.01	0.10	0.01	0.11	0.00	0.06
Village	0.21	0.41	0.24	0.43	0.22	0.41	0.16	0.37
Town	0.36	0.48	0.33	0.47	0.33	0.47	0.43	0.50
Suburb	0.10	0.31	0.13	0.33	0.09	0.29	0.09	0.29
City	0.31	0.46	0.29	0.46	0.33	0.47	0.31	0.46
No response	0.00	0.07	0.00	0.06	0.01	0.10	0.00	0.00
Urban								
urban	0.46	0.81	0.45	0.74	0.53	1.07	0.40	0.49
Observations	902		302		300		300	

The table reports descriptives for different samples. Column (1) reports summary statistics for the entire sample. In Column (2) to (4) report statistics for the three groups (Main Treatment, Control Context, Control Self-interest) separately. The sample is representative for the adult German population in age groups, income, gender, the percentage of participants living in the eastern part of Germany and the graduate population (percentage of individuals with a degree from an institution of higher education). left-right denotes the political self-classification on a left-right political spectrum.

Table 2.A.2. Average Treatment Effects

	Control Self-interest		Control Context	
	(1)	(2)	(3)	(4)
$treat_{self-interest}$	0.239	0.395		
	(2.242)	(2.197)		
$treat_{context}$			1.129	1.608
			(2.274)	(2.193)
Constant	54.46***	32.69***	53.57***	46.36***
	(1.549)	(10.13)	(1.595)	(9.422)
Controls	No	Yes	No	Yes
R^2	0.0000189	0.106	0.000411	0.126
R^2_a	-0.00165	0.0723	-0.00126	0.0932
N	602	602	602	602

Notes: This table reports OLS estimates of participants' *misogyny belief* on treatment. Columns (1) and (2) report the regression results for the control group *Control Self-interest*. Thus, $treat_{self-interest}$ is a dummy variable that equals 1 if a participant was in *Main Treatment* and 0 if a participant was in *Control Self-interest*. Columns (3) and (4) report our findings for the control group *Control Context*. Thus, $treat_{context}$ is a dummy variable that equals 1 if a participant was in *Main Treatment* and 0 if a participant was in *Control Context*. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.3. *Main Treatment*: Correlation - Donation and *misogyny belief*

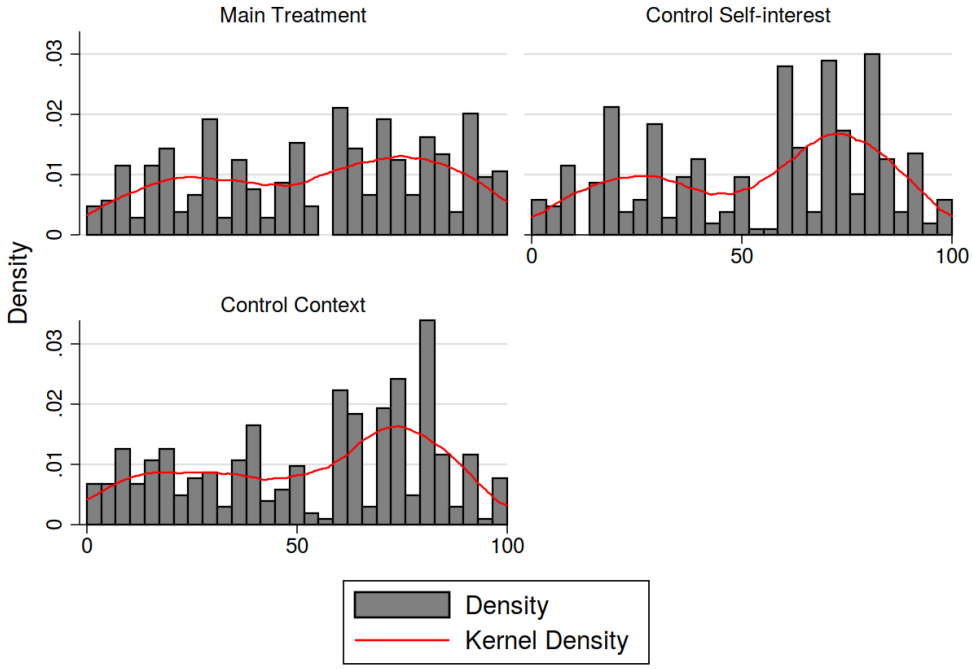
	(1)	(2)
misogyny belief	-0.211***	-0.146***
	(0.0389)	(0.0393)
Constant	39.13***	23.07*
	(2.392)	(13.47)
Controls	No	Yes
R^2	0.0893	0.276
R^2_a	0.0863	0.219
N	302	302

The table reports the correlation between donation decision and the *misogyny belief* in *Main Treatment*. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether participant live in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.4. *Control Context*: Correlation - Donation and *misogyny belief*

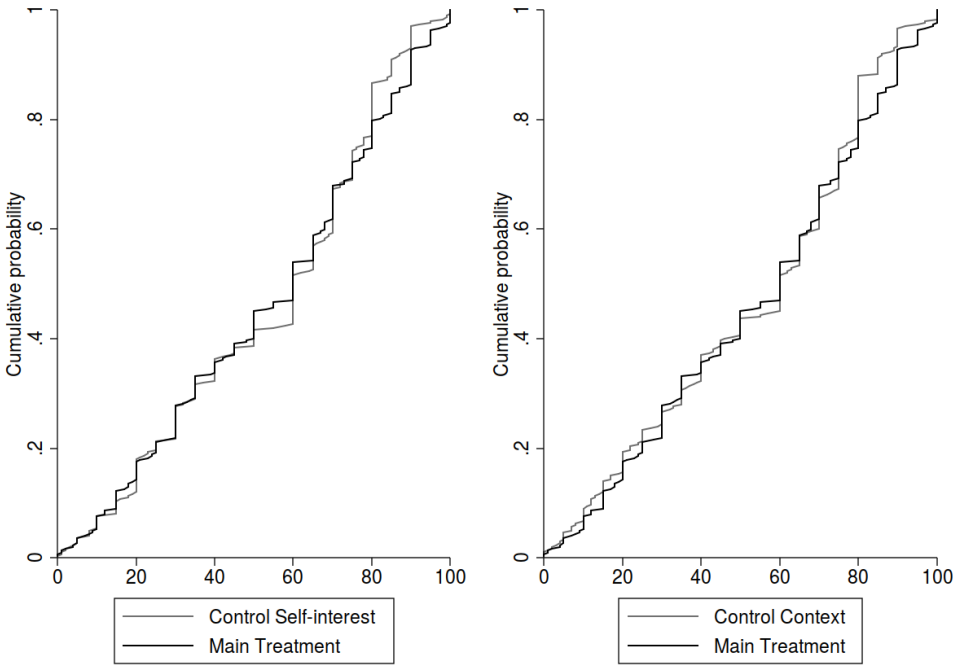
	(1)	(2)
misogyny belief	-0.0500	-0.0410
	(0.0371)	(0.0397)
Constant	34.85***	31.73***
	(2.235)	(7.900)
Controls	No	Yes
R^2	0.00605	0.0706
R^2_a	0.00272	-0.00316
N	300	300

This table reports the correlation between donation decision and the *misogyny belief* in *Control Context*. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether participant live in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$



The figure shows the distribution of *misogyny beliefs* in the three experimental conditions. Participants were asked to answer the following question: Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy? Thus, *misogyny beliefs* lies between 0 and 100. The top left histogram displays the answers in *Main Treatment*. The two other histograms report the results for the two control groups, *Control Self-interest* and *Control Context*. The red line is the kernel density estimation of the density function.

Figure 2.A.1. Distribution of *misogyny beliefs* in the experimental conditions



The figure shows the Cumulative distribution function (CDF) of *misogyny beliefs* in the three experimental conditions. Participants were asked to answer the following question: Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy? . Thus, *misogyny beliefs* lies between 0 and 100. The left hand side picture puts the CDFs of the *Control Self-interest* and *Main Treatment* into perspective. The right hand side picture compares the CDFs of the *Control Context* and the *Main Treatment*.

Figure 2.A.2. Cumulative distribution of *misogyny beliefs* in the experimental conditions

Table 2.A.5. Probit - *misogyny belief* above 3rd Quartile

	(1) larger_75
Age	0.00996***
	(3.62)
Political self-classification	0.0684**
	(1.97)
Income	-0.0484
	(-1.14)
Gender	-0.0864
	(-0.93)
East-Germany	0.0354
	(0.29)
University degree	-0.247*
	(-1.93)
Urban	0.0299
	(0.53)
Christian denomination?	0.00602
	(0.06)
Migration background	-0.0185
	(-0.33)
Constant	-1.290***
	(-5.98)
N	902
R^2_p	0.0241
χ^2	28.16

The table reports the results of our probit regression. The dependent variable *larger_75* is a dummy that equals one if a participant's *misogyny belief* is above the third quartile (*misogyny belief* above 78) and zero if the belief is below that number. Robust Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.6. Quantile Regressions - Control Self-interest

	ATE	Quantile Regression				
	(1)	(2) 15th	(3) 25th	(4) 50th	(5) 75th	(6) 85th
$treat_{self-interest}$	0.395	2.750	4.158	0.135	2.218	4.752**
	(2.197)	(2.022)	(2.582)	(3.053)	(1.799)	(1.995)
Constant	32.69***	-7.667*	-4.399	17.59	74.13***	68.78***
	(10.13)	(4.031)	(4.010)	(35.77)	(3.698)	(7.012)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.106					
R^2_a	0.0723					
N	602	602	602	602	602	602

Notes: This table reports quantile estimates of participants' *misogyny belief* on treatment. Column (1) reports the Average Treatment Effect. Columns (2) to (8) report the regression results for different quantiles. The respective percentiles are the 15th percentile (column (2)), the 25th percentile (column (3)), the second quartile (column (4)), the third quartile (column(5)) and the 85th percentile (column(6)). In all regressions *Control Self-interest* is the comparison group. Thus, $treat_{self-interest}$ is a dummy variable that equals one if a participant was in *Main Treatment* and zero if a participant was in *Control Self-interest*. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.7. Quantile Regressions - Control Context

	ATE	Quantile Regression				
	(1)	(2)	(3)	(4)	(5)	(6)
		15th	25th	50th	75th	85th
$treat_{context}$	1.608	3.333	-1.667	2.319	2.500	4.421***
	(2.193)	(2.717)	(2.236)	(2.927)	(1.666)	(1.524)
Constant	46.36***	17.33***	17.33	47.40***	77.38***	74.73***
	(9.422)	(5.941)	(13.69)	(8.463)	(10.56)	(5.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.126					
R^2_a	0.0932					
N	602	602	602	602	602	602

Notes: This table reports quantile estimates of participants' *misogyny belief* on treatment. Column (1) reports the Average Treatment Effect. Columns (2) to (8) report the regression results for different quantiles. The respective percentiles are the 15th percentile (column (2)), the 25th percentile (column (3)), the second quartile (column (4)), the third quartile (column(5)) and the 85th percentile (column(6)). In all regressions *Control Context* is the comparison group. Thus, $treat_{context}$ is a dummy variable that equals one if a participant was in *Main Treatment* and zero if a participant was in *Control Context*. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.10. *Main Treatment*: Behavior on Decision Page

	Clicks		Time spent	
	(1)	(2)	(3)	(4)
belief_first	1.902	2.019	5.143	5.748
	(1.357)	(1.373)	(6.695)	(6.904)
Constant	10.84***	19.31**	83.73***	83.78*
	(0.953)	(8.759)	(4.703)	(44.06)
Controls	No	Yes	No	Yes
R^2	0.00650	0.119	0.00196	0.0807
R^2_a	0.00319	0.0501	-0.00136	0.00818
N	302	302	302	302

Notes: This table reports the effects that the display order of the two questions on the decision page had on the behavior of the participants. We focus on the effects on participants from the *Main Treatment*. We look at two measures: the time participants spent on the decision page (Time spent) and the amount of clicks on the decision page (clicks). *belief_first* equals 0 if the respective donation decision was displayed above the *misogyny belief* question. Columns (1) and (2) report how the order affected the clicks on the decision page. Columns (3) and (4) report our findings for the outcome time spent on the decision page. Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.8. Treatment Effects Order Subsamples - Control Self-interest

	Whole Sample		Belief second (order=0)		Belief first (order=1)	
	(1)	(2)	(3)	(4)	(5)	(6)
$treat_{self-interest}$	4.711	5.242*	4.711	4.753	-4.135	-4.943
	(3.210)	(3.134)	(3.115)	(3.089)	(3.210)	(3.166)
belief_first	2.697	2.894				
	(3.176)	(3.082)				
$treat_{self-interest}$ × belief_first	-8.845**	-9.655**				
	(4.480)	(4.374)				
Constant	53.02***	29.83***	53.02***	32.41**	55.72***	27.28**
	(2.320)	(8.745)	(2.251)	(12.72)	(2.229)	(11.76)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.00750	0.115	0.00780	0.147	0.00538	0.149
R^2_a	0.00252	0.0781	0.00439	0.0811	0.00214	0.0838
N	602	602	293	293	309	309

Notes: This table reports OLS estimates of participants' *misogyny belief* on treatment for the whole sample (Column (1) and (2), the subsample for which the respective donation decision was displayed first (Column (3) and (4)) and the subsample for which the *misogyny belief* question was shown first (Column (5) and (6)). In all regressions *Control Self-interest* is the comparison group. Thus, $treat_{self-interest}$ is a dummy variable that equals 1 if a participant was in *Main Treatment* and 0 if a participant was in *Control Context*. In columns (1) and (2) we added an interaction effect ($treat_{self-interest} \times belief\_first$). The dummy equals 1 if an individual is in *Main Treatment* and *belief_first* == 1 (*misogyny belief* displayed above donation decision). Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.9. Treatment Effects Order Subsamples - Control Context

	Whole Sample		Belief second (order=0)		Belief first (order=1)	
	(1)	(2)	(3)	(4)	(5)	(6)
$treat_{context}$	5.539*	6.444**	5.539*	6.359**	-3.692	-3.365
	(3.119)	(3.010)	(3.085)	(3.051)	(3.354)	(3.185)
belief_first==1	3.083	2.905				
	(3.232)	(3.103)				
$treat_{context}$ × belief_first==1	-9.231**	-10.00**				
	(4.551)	(4.376)				
Constant	52.19***	43.82***	52.19***	37.64***	55.28***	60.76***
	(2.160)	(9.498)	(2.137)	(12.27)	(2.434)	(14.79)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.00803	0.136	0.0101	0.144	0.00429	0.217
R^2_a	0.00305	0.0996	0.00694	0.0808	0.000750	0.150
N	602	602	319	319	283	283

Notes: This table reports OLS estimates of participants' *misogyny belief* on treatment for the whole sample (Column (1) and (2), the subsample for which the respective donation decision was displayed first (Column (3) and (4)) and the subsample for which the *misogyny belief* question was shown first (Column (5) and (6)). In all regressions *Control Context* is the comparison group. Thus, $treat_{context}$ is a dummy variable that equals 1 if a participant was in *Main Treatment* and 0 if a participant was in *Control Context*. In columns (1) and (2) we added an interaction effect ($treat_{context} \times belief\_first == 1$). The dummy equals 1 if an individual is in *Main Treatment* and $belief\_first == 1$ (*misogyny belief* displayed above donation decision).
Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area.
* $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Table 2.A.11. Order Subsamples in *Main Treatment*: Correlation - donation and *misogyny belief*

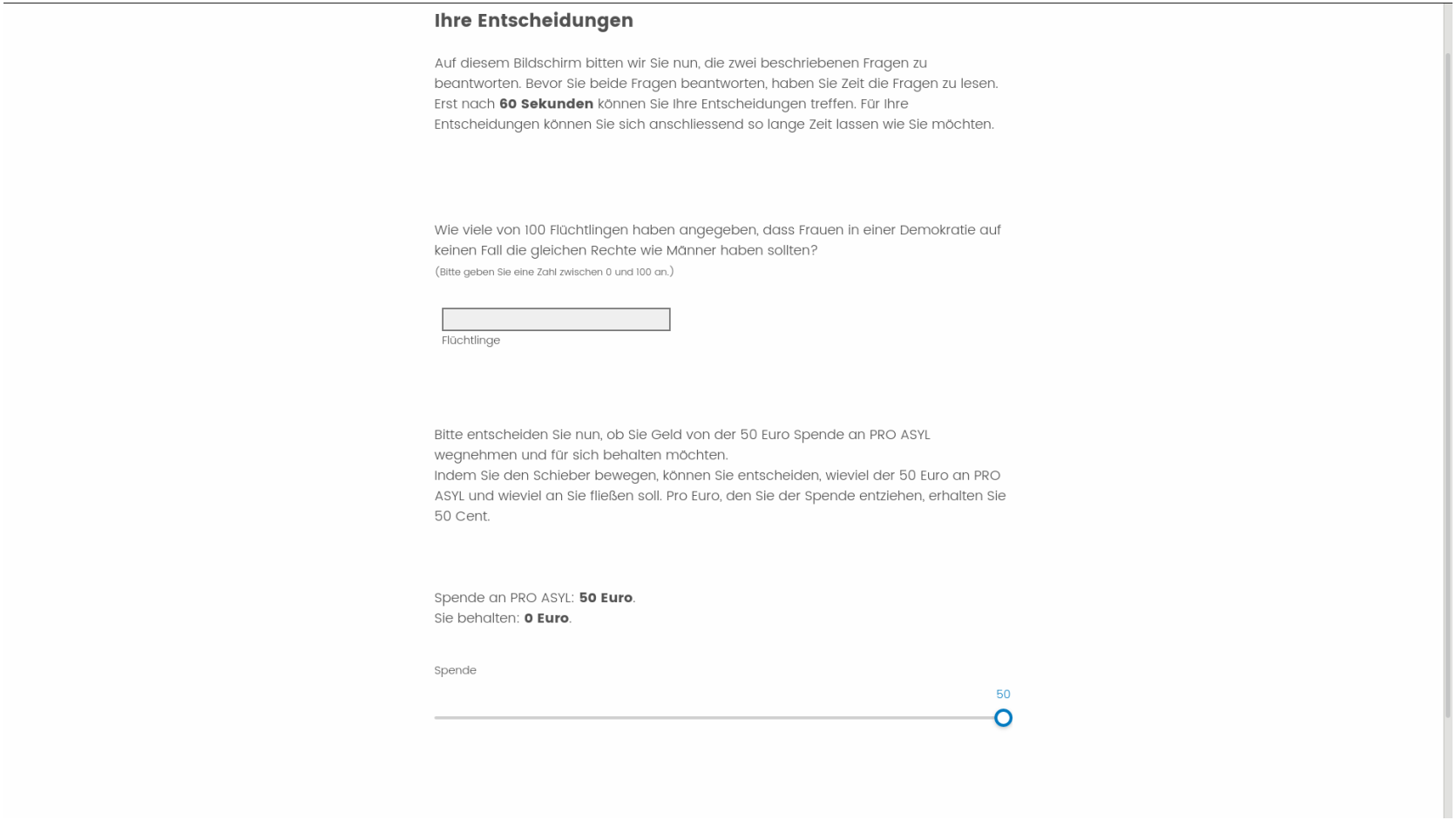
	Belief second (belief_first=0)		Belief first (belief_first=1)	
	(1)	(2)	(3)	(4)
misogyny belief	-0.274***	-0.184***	-0.150***	-0.114*
	(0.0573)	(0.0665)	(0.0533)	(0.0607)
Constant	42.26***	16.77	36.51***	13.75
	(3.657)	(25.78)	(3.149)	(15.97)
Controls	No	Yes	No	Yes
R^2	0.131	0.476	0.0514	0.420
R^2_a	0.125	0.282	0.0450	0.197
N	153	153	149	149

Notes: This table reports the correlation between donation decision and the *misogyny belief* in *Main Treatment* for two subsamples. In columns (1) and (2) we report the correlation for those participants who had the donation decision displayed above the *misogyny belief* (*belief_first* $==$ 0). The two remaining columns ((3) and (4)) report our findings for participants in belief_first=1.Robust Standard errors in parentheses. Controls are dummies for age intervals (18-25, 26-45, 46-65, >65), gender, income groups, education, area of residence (East or West Germany), political self-classification on a left-right spectrum, migration background, Christian denomination and whether a participant lives in a urban area. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

Appendix 2.B Screenshots

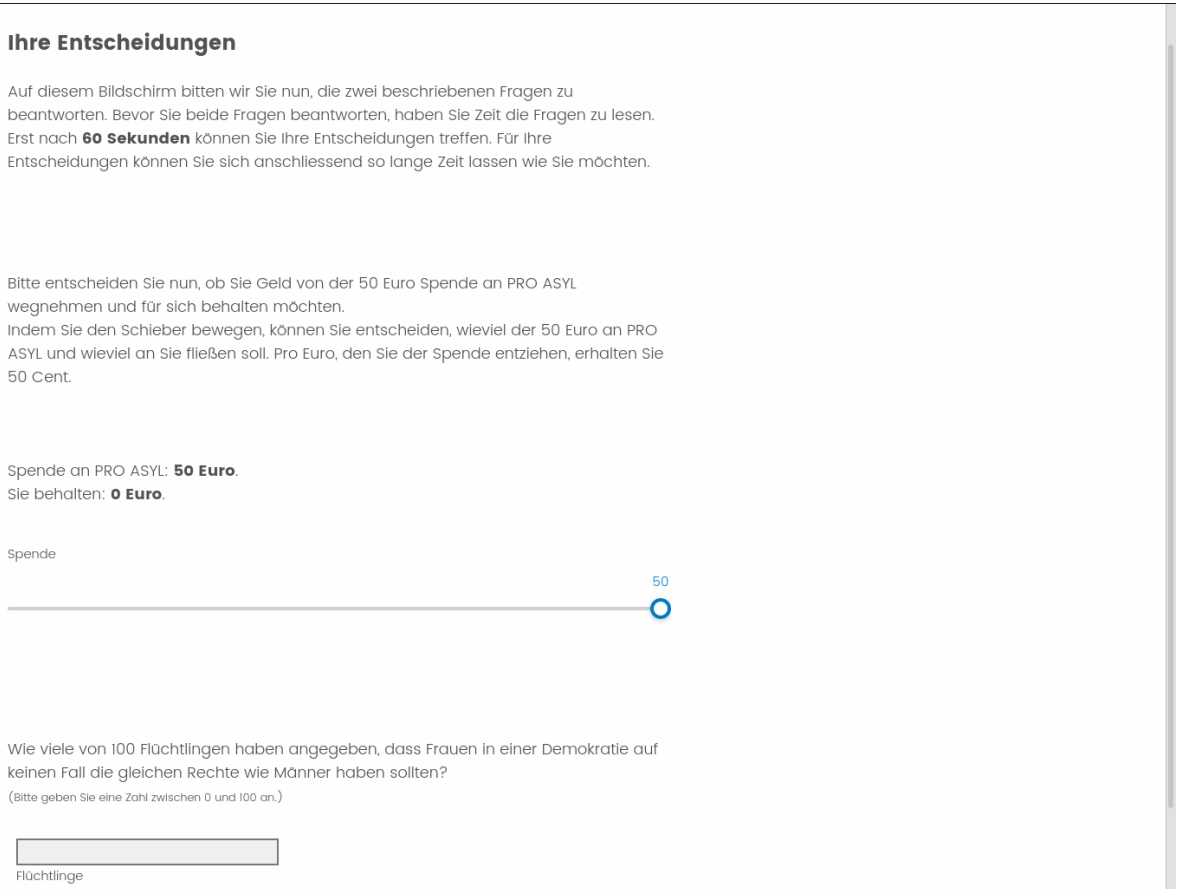
The picture shows the attention check screen. To ensure that participants read the instructions and questions carefully, they had to pass an attention check. On the attention check screen participants were told which answers they had to give to a seemingly irrelevant question. The answers participants had to give were highlighted with bold letters. Participants were only allowed to complete the survey if they passed the attention check.

Figure 2.B.1. Attention Check (Screenshot)



The picture shows the decision page in *Main Treatment*. On the decision page the two questions were randomized. This picture shows the case were the *misogyny belief* question was displayed above the donation decision ($belief\_first = 1$).

Figure 2.B.2. Order on decision page: Belief first (Screenshot)



The picture shows the decision page in *Main Treatment*. On the decision page the two questions were randomized. This picture shows the case were the donation decision was displayed above the *misogyny belief* question (*belief_first* $= 0$).

Figure 2.B.3. Order on decision page: Donation first (Screenshot)

Appendix 2.C Instructions online survey experiment

Note: Translated into English.

Page: Welcome
All your Answers will be anonymised. This means, there exists no possibility for us to lead back your answers to you.

Page: Attention Check
This question concerns the following problem: Often participants click through surveys like this one without reading the instructions carefully. As a consequence, a many of the participants' answers are of no use to us. To show us that you are reading the questions carefully, please give the following answers to the question below: "**Very interested**" and "**I've never heard of it**".

How interested are you in Game of Thrones?

a. Very interested
b. A bit interested
c. Not interested at all
d. I've never heard of it

Page: Demographics
Please answer the following questions:

- Year of birth?
- What is your state of residence?
- Sex?
  - Male
  - Female
- As of yet, what is your highest educational attainment?
  - Without degree
  - Apprenticeship
  - Degree from professional academy or vocational academy
  - Degree form a university of applied science
  - University degree
  - PhD
- What is your monthly household net income?
  - Below 1300 EUR

  - 1300 to 2600 EUR
  - 2600 to 3600 EUR
  - 3600 to 5000 EUR
  - More than 5000 EUR
- Which of the following categories describes your area of residence best?
  - City (more than 100.000 people)
  - Suburbs
  - Smaller City (more than 5.000 people)
  - Village
  - Farm
  - No response
- Many people use the words 'left' and 'right' to describe political convictions. Below you will find a scale that goes from 'left' to 'right'. When you think about your own political attitudes, where would you place yourself on the scale below?
  - Left O O O O O O O Right
- Were you and both of your parents born in Germany?
  - Yes
  - No
- What is your denomination?
  - Christianity
  - Islam
  - Buddhist
  - Jewish
  - Hindu
  - Different denomination
  - Without denomination
  - No response

1. Experimental Condition: Main Treatment

Page: Information about Study I

You will receive the amount promised by PureProfile. Additionally, you will have to make two decisions that might have financial consequences for you and others during this study. After you make the decisions, a computer will randomly choose one out of ten participants. If a participant is chosen, one of his or her decisions will be implemented. At the end of the study, we will inform you whether you were chosen or not.

In a nutshell: You will make two decisions. **If the computer chooses you, one of your two decisions will be implemented and disbursed**. You will receive the money from Pureprofile.

For your information: We are a group of scientists who are interested in your decisions and attitudes. We are not allowed to lie to you. This means that whenever we say that a decision has financial consequences for your or another person it is true. We will implement all payments as described in the instructions. We, along with Pureprofile, guarantee this.

Page: Information about Study II
We ask you to make the following two decisions:

Decision a)

We will donate 50 EUR to PRO ASYL. You have to decide whether you want to take some money away from the donation and keep it for yourself. For every EUR you take away from the donation, you will receive 50 cents. On their website PRO ASYL describes themselves as follows:

"*PRO ASYL advocates for the rights of refugees in Germany and in Europe. We help them to apply for asylum. We investigate human rights violations. And we campaign for an open society in which refugees receive protection.*"

Your first decision summarized: You have to decide whether you want to take money away from a 50 EUR donation to PRO ASYL. For every EUR you take away, you receive 50 cents. We guarantee that your decision will be implemented as described.

Page: Information about Study III
Decision b)

Furthermore we ask you to make an estimate. The famous research institute DIW Berlin conducted a survey with a representative sample of refugees who arrived in Germany between 2013 and 2016. **Your estimate concerns the refugees' image of women.** We want to know out of 100 refugees, how many said that women should in no case have the same rights as men in a democracy.

If your decision is correct, you will receive 4 EUR. The further your estimate is from the true value, the less money you will earn. You can click *here* to see the exact formula. Even if the formula looks complicated, the principle is relatively simple: The closer your estimate is to the true value, the more money you earn.

Page: Decision Page
NOTE: The order of the two decisions was randomized.

Please make the previously described two decisions. Before you answer the questions, you will have time to read them. Only after **60 seconds** will you be able to make your decisions. For your decisions you may take as much time as you need.

- "*Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy? (Please state a number between 0 and 100)*"
  - XXX Refugees

Please decide now whether you want to take away money from the 50 EUR donation to PRO ASYL.
By moving the slider, you can decide how much money should go to PRO ASYL and how much money should go to you. For every EUR you take away, you receive 50 cents.

- Donation ———————————————————X– 50
  - Donation to PRO ASYL: X
  - You receive: (50 - X)/2

2. Experimental Condition: Control Self-interest

Page: Information about Study I
You will receive the amount promised by PureProfile. Additionally, you will have to make two decisions that might have financial consequences for you and others during this study. After you make the decisions, a computer will randomly choose one out of ten participants. If a participant is chosen, one of his or her decisions will be implemented. At the end of the study, we will inform you whether you were chosen or not.
In a nutshell: You will make two decisions. **If the computer chooses you, one of your two decisions will be implemented and disbursed**. You will receive the money from Pureprofile.

For your information: We are a group of scientists who are interested in your decisions and attitudes. We are not allowed to lie to you. This means that whenever we say that a decision has financial consequences for your or another person it is true. We will implement all payments as described in the instructions. We, along with Pureprofile, guarantee this.

Page: Information about Study II
We ask you to make the following two decisions:

Decision a)

We will donate 50 EUR to PRO ASYL. You have to decide whether you want to take some money away from the donation and donate it to BumF (Bundesfachverband unbegleitete minderjährige Flüchtling). For every EUR you are taking away from the donation to PRO ASYL, BumF will receive one EUR. On their website PRO ASYL describes themselves as follows:

"*PRO ASYL advocates for the rights of refugees in Germany and in Europe. We help them to apply for asylum. We investigate human rights violations. And we campaign for an open society in which refugees receive protection.*"

On their website BumF describes themselves as follows:
"*Since 1998 the Association for Unaccompanied Refugee Minors (Bundesfachverband unbegleitete minderjährige Flüchtlinge: BumF) advocates for the rights of displaced children, adolescents and young adults in Germany. [..] It is our aim, that young refugees grow up without fear, marginalization or discrimination and enjoy the same rights as any other young person.*"

Your first decision summarized: You have to decide whether you want to take money away from a 50 EUR donation to PRO ASYL. For every EUR you take away, BumF receives 1 EUR. We guarantee that your decision will be implemented as described.

Page: Information about Study III
Decision b)

Furthermore we ask you to make an estimate. The famous research institute DIW Berlin conducted a survey with a representative sample of refugees that arrived in Germany between 2013 and 2016. **Your estimate concerns the refugees' image of women.** We want to know from you, how many out of 100 refugees said that women in a democracy should in no case have the same rights as men.

When your decision is correct you receive 4 EUR. The farer your estimate is away from the true value the less money you will earn. You can click *here* to see the exact formula. Even if the formula looks complicated, the principle is relatively simple: The closer your estimate is to the true value the more money you are going to earn.

Page: Decision Page
NOTE: The order of the two decisions was randomized.

Please make the previously described two decisions. Before you answer the questions, you will have time to read them. Only after **60 seconds** will you be able to make your decisions. For your decisions you may take as much time as you need.

- "*Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy? (Please state a number between 0 and 100)*"
  - XXX Refugees

Please decide now whether you want to take away money from the 50 EUR donation to PRO ASYL.
By moving the slider, you can decide how much money should go to PRO ASYL and how much money should go to BumF. For every EUR you take away, BumF receives 1 EUR.

- Donation ——————————————————————X– 50
  - Donation to PRO ASYL: X
  - Donation to BumF: 50 - X

3. Experimental Condition: Control Context

Page: Information about Study I
You will receive the amount promised by PureProfile. Additionally, you will have to make two decisions that might have financial consequences for you and others during this study. After you make the decisions, a computer will randomly choose one out of ten participants. If a participant is chosen, one of his or her decisions will be implemented. At the end of the study, we will inform you whether you were chosen or not.
In a nutshell: You will make two decisions. **If the computer chooses you, one of your two decisions will be implemented and disbursed**. You will receive the money from Pureprofile.

For your information: We are a group of scientists who are interested in your decisions and attitudes. We are not allowed to lie to you. This means that whenever we say that a decision has financial consequences for your or another person it is true. We will implement all payments as described in the instructions. We, along with Pureprofile, guarantee this.

Page: Information about Study II
We ask you to make the following two decisions:

Decision a)

We will donate 50 EUR to BUND (Bund für Umwelt und Naturschutz Deutschland). You have to decide whether you want to take some money away from the donation and keep it for yourself. For every EUR you take away from the donation, you will receive 50 cents. On their website BUND describes themselves as follows:

""*BUND advocates for the protection of our nature and environment - so that the earth is for all that live on it habitable*"."

Your first decision summarized: You have to decide whether you want to take money away from a 50 EUR donation to BUND. For every EUR you take away, you receive 50 cents. We guarantee that your decision will be implemented as described.

Page: Information about Study III
Decision b)

Furthermore we ask you to make an estimate. The famous research institute DIW Berlin conducted a survey with a representative sample of refugees that arrived in Germany between 2013 and 2016. **Your estimate concerns the refugees' image of women.** We want to know from you, how many out of 100 refugees said that women in a democracy should in no case have the same rights as men.

When your decision is correct you receive 4 EUR. The farer your estimate is away from the true value the less money you will earn. You can click *here* to see the exact formula. Even if the formula looks complicated, the principle is relatively simple: The closer your estimate is to the true value the more money you are going to earn.

Page: Decision Page
NOTE: The order of the two decisions was randomized.
Please make the previously described two decisions. Before you answer the questions, you will have time to read them. Only after **60 seconds** will you be able to make your decisions. For your decisions you may take as much time as you need.

- "*Out of 100 refugees, how many stated that women should in no case have the same rights as men in a democracy? (Please state a number between 0 and 100)*"
  - XXX Refugees

Please decide now whether you want to take away money from the 50 EUR donation to BUND.
By moving the slider, you can decide how much money should go to BUND and how much money should go to you. For every EUR you take away, you receive 50 cents.

- Donation ———————————————————————X– 50
  - Donation to BUND: X
  - You receive: (50 - X)/2

References

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2018. "Immigration and redistribution." Techreport. National Bureau of Economic Research. [66]

Alesina, Alberto, Elie Murard, and Hillel Rapoport. 2019. "Immigration and preferences for redistribution in Europe." Techreport. National Bureau of Economic Research. [66]

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99 (1): 544–55. [65]

Arrow, Kenneth et al. 1973. "The theory of discrimination." *Discrimination in labor markets* 3 (10): 3–33. [70]

Barone, Guglielmo, Alessio D'Ignazio, Guido de Blasio, and Paolo Naticchioni. 2016. "Mr. Rossi, Mr. Hu and politics. The role of immigration in shaping natives' voting behavior." *Journal of Public Economics* 136: 1–13. [66]

Bénabou, Rolad, Armin Falk, and Jean Tirole. 2019. "Narratives, Imperatives, and Moral Persuasion." [65, 86]

Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." *Quarterly journal of economics* 117 (3): 871–915. [69]

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American economic review* 96 (5): 1652–1678. [65]

Bénabou, Roland, and Jean Tirole. 2011. "Identity, morals, and taboos: Beliefs as assets." *Quarterly Journal of Economics* 126 (2): 805–855. [65]

Bénabou, Roland, and Jean Tirole. 2016. "Mindful economics: The production, consumption, and value of beliefs." *Journal of Economic Perspectives* 30 (3): 141–64. [69]

Bodner, Ronit, and Drazen Prelec. 2003. "Self-signaling and diagnostic utility in everyday decision making." *psychology of economic decisions* 1 (105): 26. [65]

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131 (4): 1753–1794. [69, 70]

Bund für Umwelt und Naturschutz Deutschland. 2020. "About us." https://www.bund.net/ueber-uns/. Online; accessed 20 March 2020. [74]

Bundesfachverband unbegleitete minderjährige Flüchtlinge. 2020. "About us." https://b-umf.de/en/. Online; accessed 20 March 2020. [74]

Chen, Zhuoqiong Charlie, and Tobias Gesche. 2017. "Persistent bias in advice-giving." *University of Zurich, Department of Economics, Working Paper*, (228): [69]

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics* 22 (2): 369–395. [69]

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. [69]

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism." *American Economic Review* 105 (11): 3416–42. [65, 69]

Dustmann, Christian, Kristine Vasiljeva, and Anna Piil Damm. 2019. "Refugee migration and electoral outcomes." *Review of Economic Studies* 86 (5): 2035–2091. [66]

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. [69]

Epley, Nicholas, and Thomas Gilovich. 2016. "The mechanics of motivated reasoning." *Journal of Economic perspectives* 30 (3): 133–40. [69]

Exley, Christine L. 2016. "Excusing selfishness in charitable giving: The role of risk." *Review of Economic Studies* 83 (2): 587–628. [65, 69]

Exley, Christine L. 2020. "Using charity performance metrics as an excuse not to give." *Management Science* 66 (2): 553–563. [69]

Exley, Christine L, and Judd B Kessler. 2019. "Motivated errors." Techreport. National Bureau of Economic Research. [69]

Falk, Armin. 2017. "Facing Yourself-A Note on Self-image." [65]

Falk, Armin, Thomas Neuber, and Nora Szech. 2020. "Diffusion of Being Pivotal and Immoral Outcomes." *Review of Economic Studies*, [69]

Gino, Francesca, Michael I Norton, and Roberto A Weber. 2016. "Motivated Bayesians: Feeling moral while acting egoistically." *Journal of Economic Perspectives* 30 (3): 189–212. [69]

Glaeser, Edward L. 2005. "The political economy of hatred." *Quarterly Journal of Economics* 120 (1): 45–86. [70]

Gneezy, Uri, Elizabeth A Keenan, and Ayelet Gneezy. 2014. "Avoiding overhead aversion in charity." *Science* 346 (6209): 632–635. [69]

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the self." *Games and Economic Behavior* 120: 311–324. [69]

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information avoidance." *Journal of Economic Literature* 55 (1): 96–135. [69]

Grossman, Zachary, and Joël J Van der Weele. 2017. "Self-image and willful ignorance in social decisions." *Journal of the European Economic Association* 15 (1): 173–217. [65, 69]

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. [65, 69]

Halla, Martin, Alexander F Wagner, and Josef Zweimüller. 2017. "Immigration and voting for the far right." *Journal of the European Economic Association* 15 (6): 1341–1385. [66]

Heuser, Sven, and Lasse Simon Stötzer. 2020. "Self-serving Attributions." [69]

Johnson, Eric J, Gerald Häubl, and Anat Keinan. 2007. "Aspects of endowment: a query theory of value construction." *Journal of experimental psychology: Learning, memory, and cognition* 33 (3): 461. [84]

Judd, Charles M, and Bernadette Park. 1993. "Definition and assessment of accuracy in social stereotypes." *Psychological review* 100 (1): 109. [70]

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American economic review* 90 (4): 1072–1091. [65, 69]

Köszegi, Botond. 2006. "Ego utility, overconfidence, and task choice." *Journal of the European Economic Association* 4 (4): 673–707. [69]

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108 (3): 480. [69]

Kundra, Ziva, and Lisa Sinclair. 1999. "Motivated reasoning with stereotypes: Activation, application, and inhibition." *Psychological Inquiry* 10 (1): 12–22. [70]

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." Techreport. National Bureau of Economic Research. [69]

Phelps, Edmund S. 1972. "The statistical theory of racism and sexism." *american economic review* 62 (4): 659–661. [70]

Pro Asyl. 2020. "What we do." https://www.proasyl.de/en/what-we-do/. Online; accessed 20 March 2020. [74]

Schneider, David J. 2005. *The psychology of stereotyping*. Guilford Press. [70]

Schneider, David J, Albert H Hastorf, and Phoebe Ellsworth. 1979. *Person perception*. Random House. [70]

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and self-deception." *Nature human behaviour* 3 (10): 1055–1061. [69]

Thaler, Michael. 2019. "The "Fake News" Effect: An Experiment on Motivated Reasoning and Trust in News." [69]

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." *American Economic Review* 110 (2): 337–61. [69]

Chapter 3

Self-serving attributions

Joint with Sven Heuser

3.1 Introduction

Individuals persistently overestimate their abilities. Overconfidence is a staggering phenomenon that occurs even in the presence of frequent feedback about abilities. Yet, the mechanisms driving persistent overconfidence in light of conflicting feedback are not obvious.[1] The literature on motivated reasoning argues that the desire to uphold a positive self-view leads individuals to assess feedback in a self-serving way. In this paper, we explore what role self-serving attributions play in the persistence of overconfident beliefs.

In most cases the feedback individuals receive about their abilities is wrapped in multi-dimensional uncertainty, i.e. there exists more than one potential reason that can explain the received information. Take for example the job application process; often individuals simply receive a 'Yes' or 'No' answer to their application. To learn from this feedback, individuals must attribute the decision to some underlying factor. For example, were they rejected due to their skills or due to some other, external factor like discrimination based on their gender? Social psychologists have emphasized the role of self-serving attributions in such situations. Self-serving bias in attributions often appears as an individual attributing success or positive feedback to some ego relevant factor (i.e. their abilities), while attributing negative feedback to external factors (Schneider, Hastorf, and Ellsworth, 1979). This allows individuals to keep a positive self-image in the light of otherwise damaging feedback. The side-effect of such an attribution process is that individuals' beliefs about external factors are distorted. In our paper, we focus on one such external

1. Economists have a strong interest in studying this phenomenon as it has been shown that overconfidence decisively affects individuals' decision making (see, e.g. Malmendier and Tate (2005), DellaVigna and Malmendier (2006), Kőszegi and Rabin (2006), Dohmen and Falk (2011), and Malmendier and Taylor (2015)).

factor: the belief about whether people are responsible for their own fate, i.e. the belief in a 'just world' (Lerner, 1980; Benabou and Tirole, 2006).
In this paper, we address whether individuals are self-serving in their attributions and the circumstances that facilitate such a behavior. Moreover, we explore the consequences of biased state of the world beliefs.
We employ a multi-day lab experiment to answer these questions. Studying self-serving attributions necessitates the generation of a context in which individuals care to uphold a positive self-view. Thus, we chose to give participants (noisy) feedback about their performance in an IQ-test. For most individuals intelligence is ego-relevant and therefore ideal for studying self-serving updating. In the first part of the experiment, all subjects answered selected questions from the IST2000R IQ-test. On the following day, we informed subjects that they were randomly placed in groups of ten and that we ranked all 10 group members according to their performance in the IQ-test. Subjects now had to state their beliefs about their rank in the group (Priors). In the next step, subjects received feedback about their performance in the IQ-test. The feedback consisted of three comparisons with randomly choosen group members. The outcome of the comparisons depended on the performance in the IQ-test and the *state of the world*. The *state of the world* is either *just* or *unjust*, stays unknown to the subjects, and is chosen at random at the beginning of the experiment. While in the just world the feedback only depends on the performance in the IQ-test, the 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 subjects are privileged, meaning they will be ranked above a BLUE subject regardless of their true rank. Analogously, BLUE subjects are discriminated against in the unjust world and will always be ranked below a RED group member. Thus, in an unjust state, the outcome of the comparison among the three randomly chosen members depends on the true rank, the state of the world, and the type of the subjects. In a just state, the outcome depends only on the true rank. After receiving the 3 signals, 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 the 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. By conditioning on the Bayesian predictions, we can compare the updating

2. Subjects known their type and also the exact distribution of types in the group. There are always 5 BLUE and 5 RED types.

behavior between the treatment and control group.
To illustrate the consequences of biased attributions towards the *state of the world*, all subjects had to complete a real-effort task. We implement a variation of the slider task by Gill and Prowse (2012) in which subjects had five minutes to pull as many sliders to a predetermined number as possible. Each subject was matched with a random person to compete for 4 EUR. The outcome of the competition depended on the number of sliders, on the subject's *type*, and the *state of the world*. While in the just world only the number of sliders determines who wins the money, in the unjust world RED types always win against BLUE types. Thus, the likelihood of being in the unjust world should affect the subjects' effort.[3] We conclude the experiment with two additional measures. First, subjects observed the feedback of a different person. They know the type of this person and also that the person lives in the same world. They have to assess the probability that this person is ranked in the upper half of her group. Second, subjects filled out a price list in which they decided between paying/receiving money and learning in which state of the world the subjects spent the experiment. If subjects decided to learn the state it is revealed at the end of the experiment, the decision to avoid or receive the information has no strategic component and differences in the willingness to pay have only motivational reasons.

To derive our hypotheses, we follow Eil and Rao (2011) by defining positive feedback as winning all three comparisons and negative feedback as losing all three comparisons.[4] Winning a comparisons is defined as being ranked higher. As a reminder, while in the just world only the true rank determines the outcome of the comparison in the unjust world RED types are always ranked above BLUE types. We argue that (potentially) privileged subjects (RED types) attribute positive feedback towards their intelligence and significantly understate the likelihood of being in the unjust world. This would mean that, compared to subjects in the control group, they on average state relative higher IQ-test performance beliefs and lower relative unjust world beliefs.[5] To sustain a positive self-image, we hypothesize that (potentially) discriminated subjects (BLUE types) attribute negative feedback disproportionally towards the external fundamental, i.e. the likelihood of being in the unjust world. The resulting distorted unjust world beliefs should affect the behavior in the effort

3. This dynamic can be consequential. Imagine a BLUE type who attributed her negative feedback in a motivated manner to the unjust world. Overstating the unjust World belief subsequently leads to less effort and therefore worse outcomes. Thus, in a job market scenario, self-serving attribution could even aggravate existing gaps between privileged and discriminated individuals.

4. In Appendix 3.D we relax this definition and argue that subjects that won 2 or more comparisons received positive feedback and subjects winning zero or one comparison received negative feedback.

5. A relative belief is the stated belief divided by the Bayesian prediction - this transformation allows us to compare the updating behavior between the treatment and control group. For details see 3.3.1.

task, social learning task, and willingness to pay task.
It is important to highlight that the 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 should be seen as work in progress. All results that depend on treatment-control comparison should especially be interpreted with uttermost caution.
Using the data at hand, we find that on average subjects do not attribute the observed feedback in a self-serving manner. RED types who received positive feedback state a 0.138 (significant at 1%) higher relative IQ-test performance belief in the treatment. However, the relative world belief does not differ between the treatment and control group, i.e. subjects do not make motivated attirbutions towards the state of the world. The observed effects for BLUE types with negative feedback go completely against our hypotheses: The subjects in the treatment state a significantly lower relative IQ-test performance belief (lower by -1.164) and a significantly lower relative unjust world belief(lower by -0.231).[6] Most of our behavior measures are founded on self-serving attributions towards the unjust world belief. Consequently, the observed behavior in the effort task, the updating in the learning about others task, and the willingness to pay to learn the true state of the world do not reflect the consequences of a biased world belief in the hypothesized way.
The preliminary results show no signs of self-serving attributions towards the state of the world. This leads us to step away from comparisons between the treatment and control group, and, instead, focus on potential alternative updating mechanisms that affect the updating process under multi-dimensional uncertainty. The following findings are discussed in Section 3.5.
In the strand of literature on updating with one-dimensional uncertainty, one common result is the asymmetric reaction to Good vs. Bad News (see, e.g. Eil and Rao (2011) and Mobius, Niederle, Niehaus, and Rosenblat (2011)). In particular, evidence indicates that people react disproportionately strongly when presented with good news than they do to comparable bad news. Due to the exogenous variation in the feedback (conditional on the rank), we can test for heterogeneous reactions based on the sign of the feedback.[7] We find that subjects indeed react differently to positive than to negative feedback, but only for the IQ-test performance belief. When receiving positive feedback, subjects respect the strength of the received signal significantly more, i.e. they are more willing to adapt their beliefs. This asymmetry in reactions is purely driven by RED types. BLUE types don't exhibit significant differ-

6. Especially in this case the small control group is far off. The average relative IQ-test performance belief in the control is larger than 2, i.e. the 15 subjects in the Control stated on average a belief double as high as what Bayes Rule would predict.

7. Further, the two-dimensional setting of our experiment can deepen the understanding of asymmetric updating by distinguishing between attributions to the noise component (noisy attribution bias) and the state of the world (external fundamental bias).

ences in their reactions.

Based on this finding, we study how the randomly given type affects the subjects' updating behavior. We observe that RED types are generally less responsive when it comes to their IQ-test performance belief, i.e. relative to the Bayesian prediction they do no adapt their IQ-test performance belief much. Taken together, we observe that in the unjust world privileged subjects (RED types) choose to brush aside negative feedback and only react to feedback that helps to boost their self-view. Interestingly, BLUE types take both negative and positive feedback equally to heart.

We are also able to test how initial overconfidence (overconfident priors) affects the interpretation of the feedback. Recent theoretical papers by Heidhues, Kőszegi, and Strack (2018) and Hestermann and Yaouanq (2020) show how an initial bias in the self-view leads to distorted beliefs about a related external fundamental even when individuals are Bayesian. In our experiments, biased beliefs can arise from multiple sources. However, we are still able to identify overconfident subjects using the elicited priors and check whether these subjects reacted differently towards the feedback relative to the Bayesian prediction. We find that, when it comes to their abilities, overconfident subjects react significantly less to the strength of the feedback, i.e. they are reluctant to change their distorted IQ-test performance belief. Further, by comparing subjects from the treatment group that received the same feedback and have an identical IQ-Score, we show that an overconfident self-assessment leads to a different view on the world. Indeed, overconfident RED Types have a -11.12 points lower unjust World belief than subjects with an approximately correct prior.

Overall, we show that the updating is affected by the direction of the feedback, the randomly assigned type, and initial overconfidence. These findings highlight the complexity of updating in an environment with more than one dimension of uncertainty.

Our study relates to multiple strands of literature in economics and social psychology, and more specifically to the literature on motivated reasoning and belief updating, attribution bias and overconfidence.

There is a long standing strand of literature that studies how motives affect our beliefs (see, e.g. Kunda (1990) and Epley and Gilovich (2016)). Of special interest for this paper are studies that explore how the desire to uphold a positive self-view can explain the existence and persistence of overconfidence (see, e.g. Bénabou and Ti-

role (2002), Köszegi (2006), and Sharot, Korn, and Dolan (2011).[8,9] In this context, the research on short-term updating of beliefs about an ego-relevant characteristic in the presence of uncertainty is closely related to our paper. Prominent findings from these literature are: individuals updating behavior is conservative and asymmetric, putting more weight on positive than on negative information(see, e.g.Eil and Rao (2011), Mobius et al. (2011), Barron (2016), Coutts (2019), and Schwardmann and Van der Weele (2019)).[10] Several explanations for overconfident beliefs in the light of feedback have been proposed, Zimmermann (2020) highlights the importance of memory. Additional explanations are motivational decision errors (Exley and Kessler, 2019) and information avoidance (see, e.g. Dana, Weber, and Kuang (2007) and Golman, Hagmann, and Loewenstein (2017)). We differ from these papers in two important ways: First, we add a second dimension of uncertainty that is traditionally absent from the literature, and second, we focus on self-serving attributions as the mechanism. 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 empirical and do not allow for a causal identification of the attribution bias.
In recent years self-serving attributions found their way into economic literature, with several studies highlighting the consequences of self-serving attributions in the field of CEO and trading behavior, and financial markets (see, e.g. Daniel, Hirsh-

8. Another important context in which implications of motivated reasoning have been studied is moral behavior. Research showed that individuals distort their beliefs about how other people behave (see, e.g. Di Tella, Perez-Truglia, Babino, and Sigman (2015) and Falk, Neuber, and Szech (2020)), charity performance score (see, e.g. Gneezy, Keenan, and Gneezy (2014) and Exley (2020)), risk and ambiguity preferences (see, e.g.Haisley and Weber (2010) and Exley (2016), preferences over fairness (see, e.g. Konow (2000) and Dana, Weber, and Kuang (2007)), product quality (see, e.g. Chen and Gesche (2017) and Gneezy, Saccardo, Serra-Garcia, and Veldhuizen (2020)), and discriminated outgroups (see Stötzer (2020)) to rationalize selfish actions

9. Several underlying reasons for the existence of overconfident beliefs have been suggested. Köszegi (2006) and Brunnermeier and Parker (2005) argue that people derive utility from being optimistic about themselves. Bénabou and Tirole (2002) argue that optimistic beliefs can be beneficial to personal motivation. Yet, another strand of literature promotes strategic signaling as a motive for optimistic beliefs (see, e.g. Burks, Carpenter, Goette, and Rustichini (2013) and Schwardmann and Van der Weele (2019)).

10. However, asymmetric updating is far from a robust finding: the studies by Barron (2016) and Coutts (2019) find only weak or no evidence for asymmetric updating.

leifer, 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 individuals decision making, they lack a causal identification. 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. They find that participants are self-serving in their updating behavior, not only about their contribution to the team effort but also about the teammate. The usage in our study of a world state as the external fundamental as opposed to teammate's performance both offers the ability to understand attribution in a broader context and allows us to identify how an individual's attribution differs depending on the direction of influence of this external factor.
The remainder of the paper is structured as follows. We first introduce the experimental design and procedure. Section 3.3 derives our hypotheses and outlines the empirical strategy. Section 3.4 presents the results. Section 3.5 includes discussions about alternative mechanisms and the role of initial overconfidence. Section 3.6 concludes.

3.2 Experimental Design

3.2.1 Design

A causal study of self-serving attributions in response to feedback requires an environment consisting of (i) a situation in which individuals are motivated to distort beliefs, (ii) uncertainty about the underlying reason for the feedback such that individuals can make attributions, (iii) exogenous variation in the received feedback conditional on the true ability, and (iv) a control condition in which the motives are erased. Our design accommodates all of these features.
In the experiment, participants take an IQ test. The treatment group subsequently receives feedback about their performance on the test. This feedback consists of a subject's ranking in a pairwise comparison with three other randomly chosen participants. A control group receives feedback about some other participant's performance.
Feedback depends not only on the actual performance of a subject, but also on a randomly-generated state of the world as well as the subject's type. In particular, subjects are assigned to a binary type, RED or BLUE, and the state of the world can be just or unjust. In a just world, feedback only depends on true performance. However, in an unjust world, red types will always be ranked above blue types and vice versa. We elicit subjects' beliefs about their relative performance on the test and the state of the world after they received the feedback.
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 first environmental requirement. The control test removes concerns about self-image, thereby satisfying the last requirement. The introduction of a probabilistic state of the world that crucially determines feedback implies that subjects can attribute their feedback to either their true performance or to the state of the world, or some combination of the two, thereby satisfying the third requirement. Lastly, the structure of feedback allows for exogenous variation conditional on true performance by comparing each subject with three other randomly drawn participants. This implies that individuals with the same type and performance on the IQ test could receive completely different feedback. Altogether,

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

Figure 3.1. Different Conditions in the Treatment

the experiment has a 2x2x2 (Treatment x Type x Feedback) between-subjects design. As shown in Figure 3.1, there are four different conditions in the Treatment. Subjects were randomly assigned to a type (RED or BLUE) which reflects whether they are potentially privileged or discriminated against. The other dimension is the type of feedback, i.e. whether a subject receives positive or negative feedback.[11] To study self-serving attributions, we focus on the conditions **A** and **B**. Only in these two cases subjects in the treatment have the opportunity to uphold or even boost a positive self-image by distorting their belief over the external fundamental.[12] In what follows, we describe the experimental procedure in more detail.

3.2.2 Procedure

Figure 3.2 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 some 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

11. 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 Appendix 3.D we 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.

12. In Section 3.3 we have a detailed discussion about the conditions relevant to our analysis.

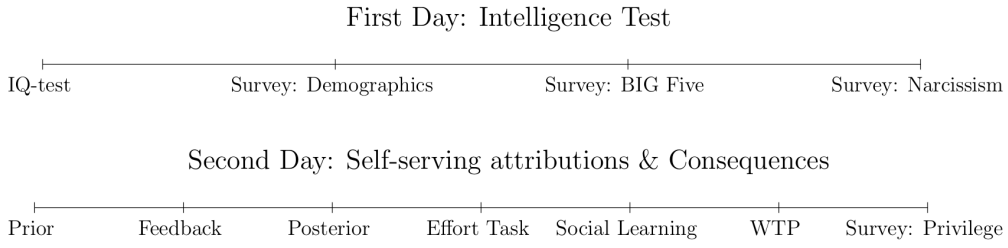


Figure 3.2. Timeline of the experiment

elicitation of the prior beliefs, (ii) a feedback stage, and (iii) the elicitation of the posterior. Subsequently, we elicited further measures to get a better understanding of the mechanisms and consequences of self-serving attributions.

Day 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 1, subjects were not told that the questions were part of an intelligence test.

Surveys. Subsequently, subjects had to fill out 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.[13]

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

SSA: Prior. After revealing that yesterday'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 answered the identical intelligence test at an earlier time.[14] Based on the subject's and the other group members' performance in the IQ-test, we calculated a ranking of the group members.[15] In the next step, we elicited subjects' beliefs about their rank in the group (Prior): First, we asked subjects to state the likelihood that they are ranked in the upper half of

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

14. A week prior to the first session, we ran a small lab experiment to construct the values for the IQ and effort task reference groups.

15. 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.

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.[16]

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.[17] In contrast to papers with only one dimension of uncertainty, the outcome of the comparisons depended on the rank of the subject, the *type* of the subject, and the *state of the world* in the following way:

- RED types were privileged in the unjust world.
- BLUE types were discriminated against in the unjust world.

Thus, if a RED type subject was compared to a BLUE type 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. If two subjects of the same type met in the unjust world the true ranks determined the outcome. In the just world, only the true ranks of the two subjects determined the outcome of the comparisons, which is to say the person with the higher IQ test score always ranked above the subject with the lower score.
As a consequence, the received feedback contained information about both an individual's rank as well as the world the experiment took place in.[18] Before observing the feedback subjects had to answer several control questions to make sure that they understood how the *type* and the *state of the world* affect the outcome of the comparisons. After passing the control questions, the subjects received their feedback. All outcomes were displayed on a single page. Subsequently, we asked subjects to repeat their feedback by asking them to state the number of comparisons they had won.

SSA: Posterior. Immediately after the feedback stage, we elicited the posterior beliefs about the intelligence and the state of the world. Subjects had to estimate

16. For details see Appendix 3.F.

17. Randomly selecting three members is crucial for our causal identification. The hereby produced noise implies that subjects with the same rank, the same type, and identical state of the world can receive different feedback. Therefore, observed asymmetries in the reaction to the received feedback cannot solely be explained by differences in demographics condition on the rank.

18. The unjust world affected the different types in an analogous but diametrical way. This helps 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.

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).[19]

Consequences. After eliciting the main outcomes (posterior beliefs), we want to highlight the consequences of motivated attributions. The (distorted) unjust world belief should affect the subjects' behavior. To show this, subjects had to participate in a real-effort task, observe the feedback of another subject, and state their willingness to pay to learn the state of the world.

Real-Effort Task 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 Next, 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 person'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 Following, we 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.[20] In Appendix 3.F we present details on how the experiment was incentivized.

Control Conditions

The control conditions 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, and that they would instead observe feedback about the performance of a different, anonymous person (Person Z).

19. Incentive compatibility was ensured using the quadratic scoring rule. Further, to rule out hedging motives between the different belief elicitation tasks, we randomly pick only one estimation per stage.

20. In the privilege questionnaire, we elicit information about the subjects' socio-economic status and his or her sexual and religious preferences.

This person was randomly chosen and participated in one of our four treatment conditions. 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.[21] At the same time, they did not receive any information about their performance in the IQ-test.
Comparisons between the updating process in the treatment and control condition allow us to causally identify if and when individuals attribute noisy feedback about their ability in a self-serving manner.

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.
All parts of the experiment were incentivized. In Appendix 3.F we describe the incentive scheme of each task in more detail.
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 Empirical Strategy and Hypotheses

Our main analysis essentially consists of two sections: First, we show if and when participants attribute feedback in a self-serving way. Subsequently, we turn to the consequences of a distorted unjust world belief. In this section, we present our empirical strategy and derive our hypotheses for both parts, the mechanism of self-serving attributions and its consequences.

21. Importantly, they did not have any motive to uphold a positive image about the other person's ability.

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 clear and testable hypotheses.

Relevant Conditions

We restrict the analysis to the feedback-type combinations that generate a motive to attribute the feedback to the state of the world in a self-serving way. We argue that people generally want to attribute positive feedback to their intelligence and negative feedback to the state of the world. Therefore, it is important to distinguish subjects based on the type of feedback they received. Concerning our design, subjects could receive either positive or negative feedback. Positive feedback is defined as being ranked first in all three comparisons, whereas negative feedback is defined as being ranked second in all three.[22] It is also important to differentiate between the two types, RED and BLUE, as they differ in their position in the unjust world. While RED type subjects are privileged, BLUE type subjects are discriminated against in the unjust world. These differentiations are relevant as not all combinations of type and feedback allow subjects to uphold a positive self-image (IQ-test performance belief) by attributing the signal to the external fundamental. For example, a RED type that receives negative feedback can not distort her belief for living in the unjust world in a way that would keep her from updating her IQ-test performance belief significantly downwards. This leads us to focus on the following two cases:

Condition A:
RED types who received positive feedback.

Condition B:
BLUE types who received negative feedback.

These conditions exclusively allow subjects to distort their beliefs about the state of the world in order to uphold a positive self-image.

22. Thus, our analysis excludes those subjects that received mixed comparisons. In Appendix 3.D we extend our definitions by including 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.

Empirical Strategy

For each of the two relevant conditions, our analysis consists of two parts. First, we investigate whether people update in a Bayesian manner. This will provide us with a first indicator as to whether subjects deviate from Bayesian updating. However, simply comparing the stated posteriors with the Bayesian predictions is insufficient. Numerous studies show that individuals often deviate from the Bayesian predictions even in the absence of motives. Therefore, in a second step, we compare updating behavior between subjects in the treatment and control groups in order to rule out other explanatory factors for deviation from Bayes Rule.

Step 1: Comparison with Bayesian prediction. Using the priors about the rank within the group of ten and the likelihood of the unjust World, we calculate the Bayesian predictions for both posterior beliefs.[23,24] To see if subjects in the two relevant treatment conditions deviate from the Bayesian prediction, we run the following regressions:

$$\textit{Posterior\_IQ}_i = \alpha + \beta_1 * \textit{Bayes\_IQ}_i + \epsilon_i \tag{3.1}$$

$$\textit{Posterior\_World}_i = \alpha + \beta_2 * \textit{Bayes\_World}_i + \epsilon_i \tag{3.2}$$

where $\textit{Posterior\_IQ}_i$ is the stated IQ-test performance belief and $\textit{Posterior\_World}_i$ the stated unjust World belief after feedback. A β_1 or 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 overreaction.
While comparing the stated posteriors with the Bayesian predictions in the treatment group provides the first indicator of attribution, it should only be seen as a signal and not be taken as conclusive evidence for self-serving motives. For example, subjects could behave in a non-Bayesian manner due to cognitive constraints triggered by the rather complex setup of our experiment.

Step 2: Comparison of Treatment and Control. To provide causal evidence on the mechanism of self-serving attributions, we compare the updating behavior in the treatment with the updating in the control. In our control conditions subjects face the same experiment but without the motive to uphold a positive self-image. As a consequence, they should share all the behavioral deviations from the Bayesian prediction except the one stemming from the interest to sustain a positive self-view. One difference is that subjects in the treatment and control group start from different priors. Subjects in the control group have no information about the

23. See Appendix 3.E for the derivations and a more formalized discussion of the self-serving attribution bias.

24. The prior belief about living in the unjust world is - as known by the subjects - 0.5. See Appendix 3.E for the derivation of the posterior belief.

performance of the randomly allocated other person, meaning they would assign uniform probability to each rank. In the treatment group, subjects form priors about their own intelligence relative to others, allowing them to have relatively informed priors about their rank. Because of the difference in priors, we can not simply compare the posterior beliefs across the groups. To make the updating behavior comparable, we construct the following two variables:

Relative IQ-test performance belief: rel_IQ
The stated posterior of being ranked in the upper half of the group relative to the Bayesian prediction:

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

Relative unjust world belief: rel_World
The stated posterior of living in the unjust world relative to the Bayesian prediction:

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

Looking at the relative posteriors erases the problem stemming from unequal priors, thereby rendering the updating behavior in the treatment and control group comparable.[25] A $rel\_IQ_i > 1$ ($rel\_World_i > 1$) implies that the subject stated a posterior belief that is higher than what Bayes' theorem would predict. To detect differences between the treatment and the control groups, we run the following regressions:

$$rel\_IQ_i \;=\; \alpha + \beta_3 * treat_i + \gamma * Controls_i + \epsilon_i \tag{3.3}$$

$$rel\_World_i \;=\; \alpha + \beta_4 * treat_i + \gamma * Controls_i + \epsilon_i \tag{3.4}$$

where $treat_i$ is a dummy indicating whether the subject was in the treatment or the control group. To study the robustness of these effects, we add control variables in two steps. First, we add a control variable for the score in the IQ-test. The score in the IQ-test helps to control for the performance of the subject and indirectly for her rank in the group.[26] Second, we add several controls to study whether certain characteristics of the subjects drive the observed results. Namely, 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. We run the two regressions for the two relevant conditions separately because the hypothesized attributions differ greatly in the two cases.

25. There exists a debate about whether starting from different priors effects updating mechanically (see, 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.

26. This is important as the received feedback is still dependent on the performance in the IQ-test.

Hypotheses Self-serving attributions

Condition 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 (3.3): 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 ($\beta_3 > 0$).
- Equation (3.4): 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 ($\beta_4 < 0$).

Condition B (BLUE + negative feedback):

Hypothesis 1.2. Subjects who are potentially discriminated against (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 (3.3): 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 ($\beta_3 > 0$).
- Equation (3.4): 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 ($\beta_4 > 0$).

3.3.2 Consequences

Above, we hypothesized that due to self-serving attributions, subjects in our two relevant cases end up with distorted beliefs about the *state of the world*. To showcase the consequences of self-serving attributions on the subjects' decision making, we implemented a real-effort task and ask subjects to make inferences about a different person. Further, we elicited the monetary willingness to pay to learn the true state of the world. The following hypotheses critically depend on self-serving attributions. In the absence of such attributions, we shouldn't expect any effects.

Real Effort Task

In a first step, we turn towards the consequences of self-serving attributions by investigating how the unjust world belief affects subjects' effort. 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 run the following OLS regression for the subjects in the two relevant cases:

$$\text{Effort}_i \;=\; \alpha + \beta_5 * \textit{Posterior\_World}_i + \epsilon_i \tag{3.5}$$

Hypothesis 2. The subjects' effort significantly sinks with the belief to live in the unjust world.

- Equation (3.5): $\beta_5 < 0$

Social Learning

Does the distorted unjust world belief affect the perception of other people? We argue that subjects in our two relevant cases aim to uphold their unjust world belief as it enables them to maintain their positive self-image. When being confronted with conflicting information, 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 choose to make strong inferences about the other person, we run the following regression:

$$\textit{Social\_learning\_rel\_IQ}_i \;=\; \alpha + \beta_6 * \textit{treat}_i + \gamma * \textit{Controls}_i + \epsilon_i \tag{3.6}$$

where $\textit{Social\_learning\_rel\_IQ}_i$ is the relative IQ-test performance belief about the person whose feedback the subjects observed.
Condition A (RED + positive feedback):

Hypothesis 3.1. To uphold a positive self-image and world belief, RED types significantly understate the probability that the other person is in the upper half of her group.

- Equation (3.6): $\beta_6 < 0$.

Condition B (BLUE + negative feedback):

Hypothesis 3.2. To secure the positive self-image and their world belief, BLUE types significantly overstate the probability that the other person is in the upper half of her group.

- Equation (3.6): $\beta_6 > 0$.

Willingness to pay

The literature 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 potentially conflicting information, 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 can test this by running the following regression with subjects from our two relevant cases:

$$Willingness\_to\_pay_i = \alpha + \beta_7 * treat_i + \gamma * Controls_i + \epsilon_i \tag{3.7}$$

Hypothesis 4. Subjects in the treatment group have a lower monetary willingness to pay to learn the true state of the world.

- Equation (3.7): $\beta_7 < 0$

Having derived testable hypotheses for self-serving attributions and potential consequences of the distorted unjust world belief, we now turn to our results.

3.4 Results

As outlined in the previous section, the upcoming results section is divided into two parts. First, we focus on updating behavior. In particular, we seek to establish whether and under what circumstances individuals attribute feedback in a self-serving manner. Subsequently, we turn to our further measures and study the consequences of self-serving attributions.

3.4.1 Self-Serving Attributions

As laid out in the previous section, we focus on the two relevant treatment conditions. For each of the two cases, we first study if the subjects' stated posteriors deviate from the Bayesian predictions in the treatment group. Afterward, we compare the updating behavior between treatment and control groups.

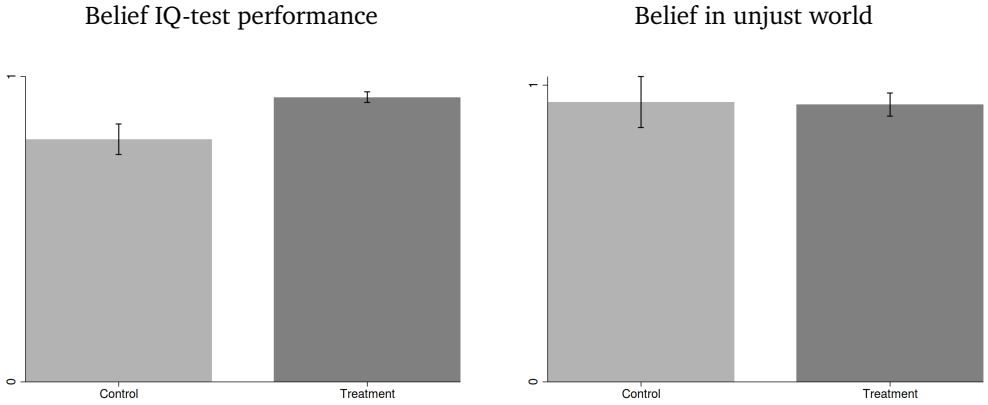
Remark: As highlighted before, we did not complete our data collection process. As a consequence, our control groups are comparably small. Concerning Condition A (RED type + positive feedback), there are 56 subjects in the treatment but only 17 subjects in the control group. Similarly, regarding Condition B (BLUE + negative feedback), there are 56 subjects in the treatment but only 15 subjects in the control group. The analysis here should then be seen as a proof of concept

as opposed to a thorough and comprehensive investigation of the research questions.

Condition A (RED + positive feedback):

Step 1: Comparison with Bayesian predictions. In a first step, we compare the stated posteriors in the treatment groups with the Bayesian predictions (see equations (1) and (2)). Table 3.A.1 reports our findings. The coefficient of the IQ-test performance prediction is 0.911 and for the unjust World prediction the coefficient is 1.170. We find that neither coefficients significantly differs from one, i.e. we cannot reject the hypothesis that the subjects behave like a Bayesian updater. However, as discussed in Section 3.3.1 and Appendix 3.E, this does not necessarily mean that subjects do not attribute the feedback in a self-serving manner.

Step 2: Comparison of Treatment and Control. In the second step, we compare the relative updating behavior between the treatment and the control groups. While we report our findings, we should always keep in mind the missing data. In Figure 3.1 we plot the mean relative posterior beliefs, which corresponds to



Note: The left picture plots the average relative posterior of the likelihood for being ranked in the upper half ($rel_{IQ} = \frac{Posterior_{upper\_half}}{Bayes_{upper\_half}}$) in the control and treatment. The right picture plots the average relative posterior of the probability for being in the unjust world ($rel_{World} = \frac{Posterior_{World\_unfair}}{Bayes_{World\_unfair}}$). The error bars indicate $+/-$ standard errors.

Figure 3.1. Treatment Effects: Condition A (RED + positive feedback)

the outcomes of regressions (3.3) and (3.4) in the case without any controls. The results of the underlying regressions with and without controls are reported in Table 3.A.3. The preliminary findings from Figure 3.1 reveal that the relative IQ-test performance belief (*rel_IQ*) is significantly higher in the treatment and does not differ for the relative unjust world belief (*rel_World*). The corresponding numbers from the regression in Table 3.A.3 column (1) confirm this observation: Focusing on the relative IQ-test performance belief, we observe that the coefficient for the

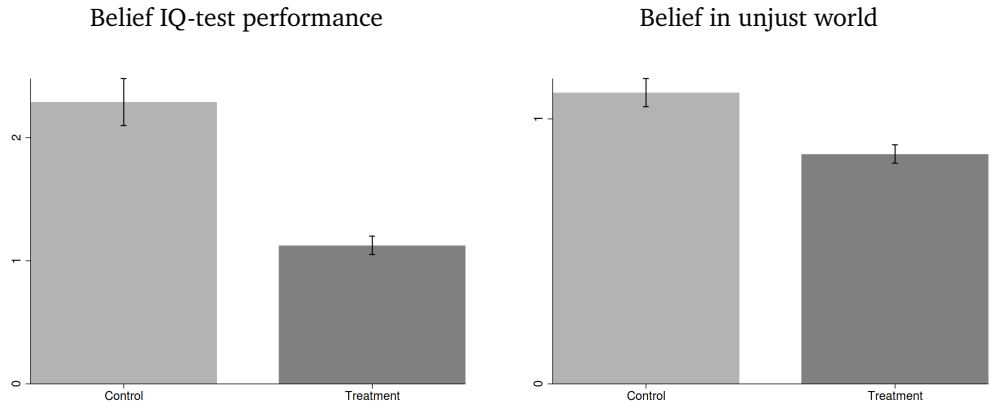
treatment dummy in our specification without controls is 0.138 (significant at 1%), showing that subjects in the treatment take more credit for the positive feedback. The mean *rel_IQ* in the Treatment is 0.933 and in the Control 0.795. Recall that our theory of self-serving attributions implies that while people take more credit for the positive feedback, they should simultaneously understate the role the world (external fundamental) played for the received feedback. Figure 3.1 and Table 3.A.3 show that subjects in the treatment do not differ in their relative updating behavior over the unjust world belief. The mean *rel_World* in the treatment is 0.935 and in the control 0.943. Thus, while subjects in our treatment give themselves disproportionate credit for the positive feedback, they do not make self-serving inferences over the external fundamental. As highlighted in the remark, at this stage of the data gathering process these findings are rather inconclusive.[27]

Condition B (BLUE + negative feedback):

Step 1: Comparison with bayesian predictions. Following the same procedure as above, we first study whether BLUE types with negative feedback deviate from the Bayesian predictions. Table 3.A.2 reports our findings. For the IQ-test performance belief the coefficient of the Bayesian prediction is 0.712, indicating that subjects are relatively unresponsive to the feedback when it comes to updating over their IQ-test performance. For the unjust World belief the coefficient is 1.006. The behavior does not deviate significantly from the Bayesian prediction.

Step 2: Comparison of Treatment and Control. Again, to rule out other mechanisms like conservatism, we compare the relative updating behavior between the treatment and control group. We report our findings in Figure 3.2 and Table 3.A.4. Figure 3.2 and Table 3.A.4 column (1) show that BLUE types with negative feedback report significantly lower relative beliefs for both the probability of being ranked in the upper half of the group (relative IQ-test performance belief) and the probability for being in the unjust world (relative World unjust belief). For the relative IQ-test performance belief the coefficient of the treatment dummy is -1.164 (significant at 1%) and for the relative unjust world belief the coefficient is -0.231 (significant at 1%). It is important to highlight that the mean *rel_IQ* in our Control is peculiar ($rel\_IQ = 2.288$) and can only be explained by the low number of subjects. As pointed out numerous times, we should not put too much

27. To get a larger comparison group and test for the robustness of our findings we also investigate what happens when we loosen our definition of positive feedback. It is important to note that our control group is still heavily underpowered. The proportions stay the same when changing the definition of positive feedback: For 3 subjects in the treatment, we have only 1 subject in the control. We run the same regressions as above, this time looking at RED type participants that won all three comparisons or won two of the three comparisons. Table 3.D.1 reports the findings for this new subject pool. We observe that neither the relative unjust World belief nor the relative IQ-test performance belief differs between Treatment and Control in this subject pool.



Note: The left picture plots the average relative posterior of the likelihood for being ranked in the upper half ($rel_{IQ} = \frac{Posterior_{upper\_half}}{Bayes_{upper\_half}}$) in the control and treatment. The right picture plots the average relative posterior of the probability for being in the unjust world ($rel_{World} = \frac{Posterior_{World\_unfair}}{Bayes_{World\_unfair}}$). The error bars indicate $+/-$ standard errors.

Figure 3.2. Treatment Effects: Condition B (BLUE + negative feedback)

emphasis on the observer results. Although we observe disproportionate updating in both dimensions, the directions of updating are opposite to what our hypothesis predicted.[28]

Taking the results for a moment at face value, we have to conclude that we see no evidence for self-serving attributions to the external *state of the world*. Neither RED types that received positive feedback nor BLUE types that received negative feedback distort their unjust World belief in a motivated manner. In Section 3.5, we will discuss factors that help us to get a more complete understanding of updating under two-dimensional uncertainty.

3.4.2 Consequences

Working with the data we have at hand, we observe that subjects' in the treatment group do not distort the unjust world belief in the hypothesized way. As all of our consequence measures are founded on 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. We do however run all the described regressions and present the results in Appendix 3.B.

28. When expanding the analysis to subjects that lost all or won only 1 comparison, we observe a similar pattern. Table 3.D.2 shows that the *rel_IQ* is still significantly smaller in the treatment group, but for the unjust World belief we no longer observe significant differences.

3.5 Discussion

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 Condition 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.

3.5.1 Alternative Mechanism

Attribution to Noise or Good vs. Bad News Effect

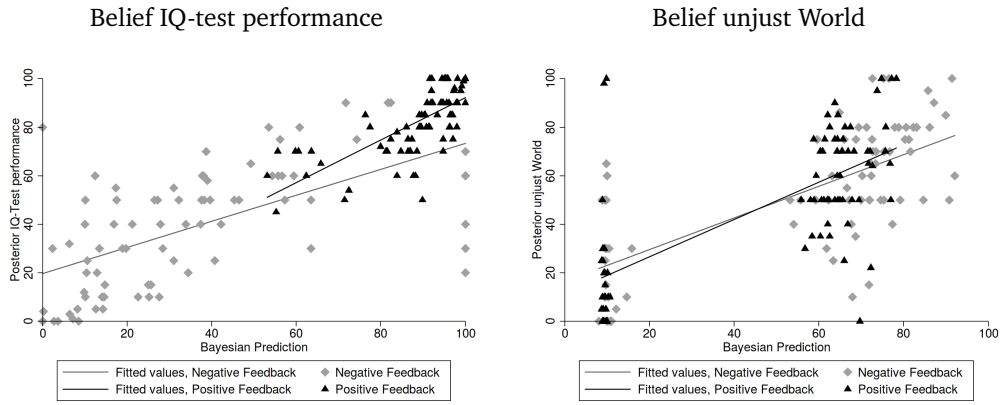
A prominent hypothesis in the literature on short term updating with one dimension of uncertainty is that individuals react more strongly to favorable news than to unfavorable news (see, e.g. Eil and Rao (2011)). When individuals are confronted with negative feedback they surmise that the noise component 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.[29,30] Figure 3.1 plots the reaction depending on whether subjects received

29. The regression are as follows:

$$\begin{aligned} \textit{Posterior\_IQ}_i &= \alpha + \beta_1 * \textit{Bayes\_IQ}_i + \beta_2 * \textit{good\_news} \\ &\quad + \beta_3 * \textit{Bayes\_IQ}_i * \textit{good\_news} + \gamma * \textit{Controls}_i + \epsilon_i \\ \textit{Posterior\_World}_i &= \alpha + \beta_1 * \textit{Bayes\_World}_i + \beta_2 * \textit{good\_news} \\ &\quad + \beta_3 * \textit{Bayes\_World}_i * \textit{good\_news} + \gamma * \textit{Controls}_i + \epsilon_i \end{aligned}$$

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.

30. 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 or 3 comparisons) and mostly bad news (won 0 or 1 comparison). See Tables 3.D.6, 3.D.7, and 3.D.8.



Note: 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.

Figure 3.1. Updating Behavior: Good News vs. Bad News

good or bad news. Different slopes indicate that subjects reacted differently to positive 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.C.1 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.C.2, 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.C.3). 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.[31] This raises the question if the randomly

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.

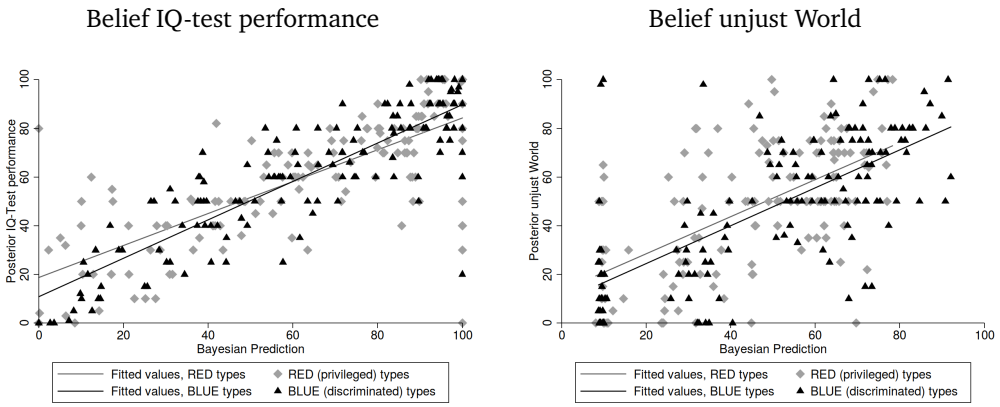
31. The effects do not change when adapting our definition of good and bad News to (mostly) positive and (mostly) negative feedback (see Table 3.D.6 for all types, Table 3.D.7 for RED types only, and Table 3.D.8 for BLUE types only).

assigned type affects the updating of the subjects in general.

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.[32,33] Figure 3.2 and Table 3.C.4 report our findings. Figure 3.2 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



Note: 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.

Figure 3.2. Updating Behavior: RED vs. BLUE types

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

$$\begin{aligned} \textit{Posterior\_IQ}_i &= \alpha + \beta_1 * \textit{Bayes\_IQ}_i + \beta_2 * \textit{type} \\ &\quad + \beta_3 * \textit{Bayes\_IQ}_i * \textit{type} + \gamma * \textit{Controls}_i + \epsilon_i \\ \textit{Posterior\_World}_i &= \alpha + \beta_1 * \textit{Bayes\_World}_i + \beta_2 * \textit{type} \\ &\quad + \beta_3 * \textit{Bayes\_World}_i * \textit{type} + \gamma * \textit{Controls}_i + \epsilon_i \end{aligned}$$

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

does not differ between types over the unjust world belief.[34] In contrast, updating behavior over the IQ-test performance belief differs significantly. As seen in the left picture in Figure 3.2, the slope of the BLUE types is steeper, indicating that BLUE participants are more responsive to the feedback. The interaction term between type and Bayesian prediction in Table 3.C.4 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.

3.5.2 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*).[35] As we seek to

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

35. 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

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.C.5 reports our findings. Focusing on the updating behavior over the IQ-test performance (Table 3.C.5 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.C.5 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.C.6 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

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.

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.

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 an additional noisy component determined by the type of the subject and the state of the world. RED types were privileged and BLUE types discriminated against in the unjust world. To learn from the feedback, subjects had to make attributions. We hypothesized that these attributions would be self-serving, meaning they would be made such as to uphold or gain a positive self-image. To provide causal evidence on self-serving attributions, we run 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 results described below must be considered with the utmost caution. Using the data we have at hand, we find no evidence of self-serving attributions. RED types in the treatment group differed significantly in their updating over the IQ-test performance, stating higher relative posteriors about the likelihood that they are ranked in the upper half of their group of ten. The relative unjust world beliefs did not differ, showcasing that RED types do not attribute positive feedback to a potentially unjust world any more than a subject in the control group does. BLUE types report significantly lower relative IQ-test performance and unjust world beliefs. This evidence contradicts self-serving attributions. All in all, we do not see that subjects' learning over the external fundamental is biased in the hypothesized way. Ergo, our consequence measures did not show significant effects in the hypothesized way.
Focusing on subjects in the treatment group only, we discuss how the sign of the feedback (Good vs. Bad News), the type of the subject and initial overconfidence affects the updating process. While RED types react more strongly to positive feedback, BLUE types' reactions do not differ based on the sign of the feedback received. Moreover, we observe that the randomly assigned type affects the updating behavior. BLUE types in general respect the signal strength more than RED types. In other words, RED types amplify the importance of positive feedback and ignore the signal

strength for other feedback. We further find that overconfident subjects are, when it comes to their IQ-test performance, less responsive to the received feedback, i.e. they are reluctant to adapt their distorted beliefs. These results show that the updating process in a setting with two-dimensional uncertainty is a multi-layered and complex problem which needs further studying.

Naturally, the first step in improving our research would be the completion of the control sessions. Having a balanced control and treatment group will provide us with the power to present conclusive evidence on self-serving attributions. Another interesting direction for future work is to adapt the here proposed design in a way that helps to distinguish between motivated attributions to the unjust world, motivated attributions to the noise component, and the influence of overconfident priors.

Appendix 3.A Aditional Tables & Figures Self-Serving Attributions

Table 3.A.1. Are Subjects Bayesian - Condition A (RED + positive feedback)

	IQ-test performance			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 (Bayesian Prediction=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
R^2_a	0.555	0.550	0.543	0.0899	0.0991	-0.00150
N	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 said to be positive when a subject won all three comparisons. In the 5th line (P-Value) we report the p-value of the relevant hypothesis test. Namely, we test whether the correlation coefficient of the Bayesian Prediction equals one. Given a p-value smaller than 0.05, we would reject the hypothesis and conclude that the coefficient does not equal one, i.e. subjects are not Bayesian in their updating.
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.2. Are Subjects Bayesian - Condition B (BLUE + negative feedback)

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	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 (Bayesian Prediction=1)	0.00194	0.00138	0.0373	0.982	0.772	0.641
R^2	0.545	0.555	0.757	0.202	0.207	0.463
R^2_a	0.537	0.538	0.602	0.187	0.177	0.121
N	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 said to be positive when a subject won all three comparisons. In the 5th line (P-Value) we report the p-value of the relevant hypothesis test. Namely, we test whether the correlation coefficient of the Bayesian Prediction equals one. Given a p-value smaller than 0.05, we would reject the hypothesis and conclude that the coefficient does not equal one, i.e. subjects are not Bayesian in their updating.
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. Treatment Effects: Condition A (RED + positive Feedback)

	Relative IQ-test performance			Relative unjust World		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Dummy	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
R^2_a	0.125	0.113	0.0262	-0.0139	-0.0232	-0.0536
N	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 said to be 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. 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. Treatment Effects: Condition B (BLUE + negative Feedback)

	Relative IQ-test performance			Relative unjust World		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Dummy	-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
R^2_a	0.388	0.380	0.491	0.115	0.107	-0.00461
N	71	71	70	71	71	70

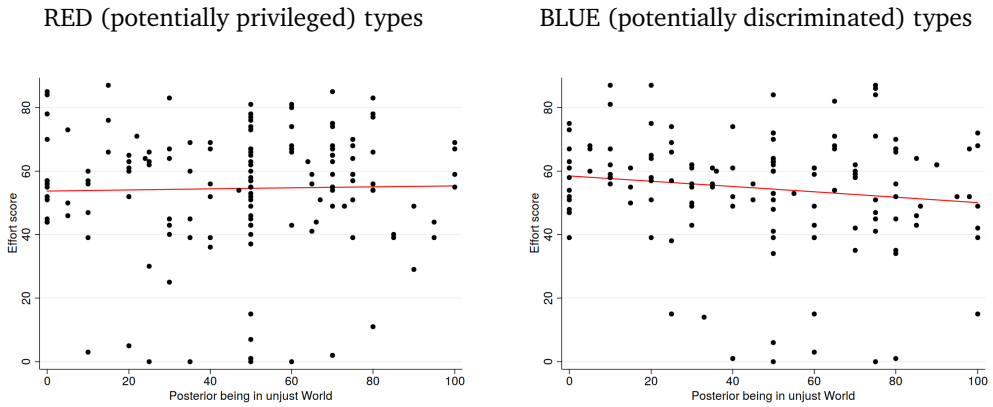
Notes: This table reports OLS estimates of subjects' relative posteriors on treatment for BLUE types that received negative feedback. Feedback is said to be negative when a subject lost all three comparisons. The treatment Dummy equals 1 if a subject received feedback about their own performance and 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. 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.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-



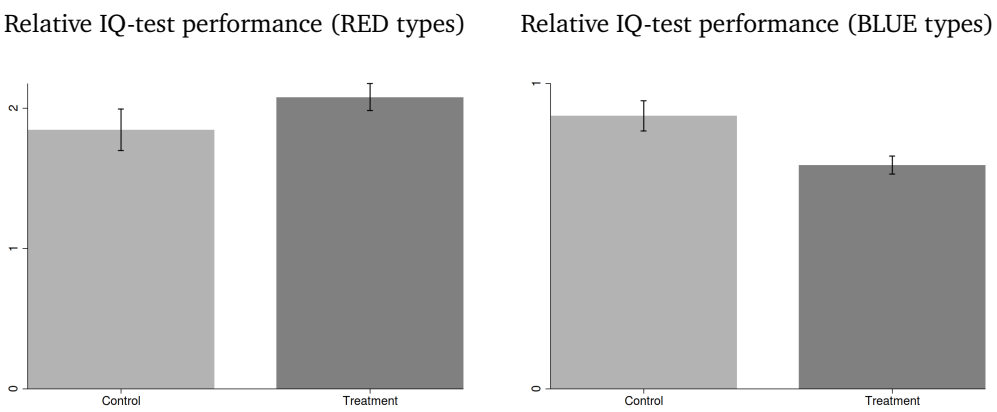
Note: The two pictures plot the correlation between subjects' unjust world Posterior belief and the amount of sliders they finished in the real effort task (effort score). The left picture shows the relation for RED (potentially privileged) types and the right hand side picture the relation for BLUE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

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

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.B.1 and 3.B.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.B.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



Note: 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.

Figure 3.B.2. Learning about Others

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.B.3 and Table 3.B.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.B.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.

Additional Tables consequences

Table 3.B.1. Correlation: Effort Score and unjust World Belief - 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
R^2_a	-0.00623	0.0567	0.150
N	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.B.2. Correlation: Effort Score and unjust World Belief - BLUE types

	(1)	(2)	(3)
Posterior unjust World	-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
R^2_a	0.0132	0.0967	0.0892
N	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.B.3. Learning about Others - Condition A (RED + positive feedback)

	(1)	(2)	(3)
Treatment Dummy	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
R^2_a	0.00612	0.00477	-0.0331
N	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.B.4. Learning about Others - Condition B (BLUE + negative feedback)

	(1)	(2)	(3)
Treatment Dummy	-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
R^2_a	0.0723	0.134	-0.0402
N	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.B.5. Willingness to pay to learn state of the world

	(1)	(2)	(3)
Treatment Dummy	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
R^2_a	-0.00622	-0.0141	-0.0184
N	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 *** $p < 0.01$

Appendix 3.C Additional Tables Discussion

Table 3.C.1. Updating Behavior: Bad News vs Good News

	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
R^2_a	0.722	0.733	0.743	0.485	0.482	0.515
N	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.C.2. Updating Behavior: 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
R^2_a	0.748	0.756	0.827	0.437	0.437	0.457
N	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.C.3. Updating Behavior: 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
R^2_a	0.746	0.748	0.750	0.523	0.521	0.527
N	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.C.4. Updating Behavior: 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
R^2	0.655	0.678	0.686	0.726	0.418	0.418	0.419	0.496
R^2_a	0.651	0.673	0.681	0.692	0.412	0.410	0.409	0.434
N	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.C.5. Updating Behavior: Initial Overconfidence

	IQ-test performance			Unjust World		
	(1)	(2)	(3)	(4)	(5)	(6)
Bayesian Prediction	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 Posterior × 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
R^2_a	0.610	0.620	0.638	0.437	0.435	0.430
N	223	223	219	223	223	219

Notes: This table reports how the subjects' response varied depending on initial overconfidence. *over* 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.C.6. Unjust World belief: Effect of Initial Overconfidence

	RED type		BLUE type	
	(1)	(2)	(3)	(4)
over	-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
R^2_a	0.342	0.313	0.484	0.449
N	120	119	103	100

Notes: This table reports the effects of initial overconfident beliefs on the unjust World belief. *over* 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 *** $p < 0.01$

Appendix 3.D Mostly positive or mostly negative feedback

In contrast to the analyses in the main part of the paper, here we relax our definition of positive and negative feedback. In the following, mostly positive feedback is defined as winning 2 or 3 comparisons, meaning a subject ranked first in the majority of comparisons, and mostly negative feedback as winning 0 or 1 comparisons.

Table 3.D.1. Treatment Effects: Condition A (RED + (mostly) positive Feedback)

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Dummy	-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
R^2_a	-0.00815	-0.0114	-0.0745	-0.000465	-0.00801	-0.0316
N	124	124	123	124	124	123

Notes: This table reports OLS estimates of subjects' relative posteriors 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 subject received feedback about their own performance and 0 if a subject observed the feedback of a random other person. Columns (1) to (3) report results of relative IQ-test performance belief and Columns (4) to (6) on the relative 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.D.2. Treatment Effects: Condition B (BLUE + (mostly) negative Feedback)

	IQ-test performance			World unjust		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Dummy	-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
R^2_a	0.322	0.315	0.339	0.00951	0.0292	0.0933
N	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 said to be (mostly) negative when a subject lost all three or won only one comparisons. 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) report results of the relative IQ-test performance belief and Columns (4) to (6) on the relative 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.D.3. Learning about Others - Condition A (RED + (mostly) positive Feedback)

	(1)	(2)	(3)
Treatment Dummy	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
R^2_a	-0.00504	-0.0123	-0.0773
N	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.D.4. Learning about Others - Condition B (BLUE + (mostly) negative Feedback)

	(1)	(2)	(3)
Treatment Dummy	-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
R^2_a	0.0268	0.0984	0.0847
N	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.D.5. Willingness to pay to learn state of the world - all participants

	(1)	(2)	(3)
Treatment Dummy	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
R^2_a	-0.00284	-0.00583	0.00576
N	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 Dummy 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.D.6. Updating Behavior: (mostly) Bad News vs. (mostly) Good News

	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
R^2_a	0.683	0.692	0.699	0.410	0.409	0.434
N	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.D.7. Updating Behavior: (mostly) Bad News 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
R^2_a	0.677	0.681	0.711	0.348	0.347	0.372
N	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.D.8. Updating Behavior: (mostly) Bad News 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
R^2_a	0.743	0.747	0.745	0.473	0.484	0.498
N	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.E Hypotheses Self-serving attributions

In this appendix we aim to establish the hypotheses from Section 3.3.1 in a more formal fashion. To do this, we first derive the Bayesian predictions and in two subsequent steps introduce the self-serving attribution bias and other behavioral deviations from the Bayesian predictions.

Bayesian prediction

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 group of ten. Let F be the feedback the subject receives.

$$Bayes\_IQ \;=\; P(\text{upper half}|F) \;=\; \frac{P(F|\text{upper half}) * P(\text{upper half})}{P(F)} \tag{3.E.1}$$

where:

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

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

The Bayesian Prediction for the unjust World Belief:

$$Bayes\_World \;=\; P(unjust|F) \;=\; \frac{P(F|unjust) * P(unjust)}{P(F)} \tag{3.E.2}$$

where:

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

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

Self-serving attributions

We argue that individuals make attributions that allow them to maintain a positive self-view. In our analysis we focus on two relevant cases:

1. RED types (privileged) with positive feedback
2. BLUE types (discriminated) with negative feedback

We argue that potentially privileged individuals who receive positive feedback attribute the feedback to their abilities instead of external factors that work to their advantage. In contrast, potentially discriminated individuals attribute their negative feedback mainly to the unfavorable state of the world. To formalize this motivated

bias, we introduce the parameter γ. Motivated individuals distort their posteriors in the following way:

Condition A. RED types with positive Feedback:

Individuals overstate the strength of the signal given that they are ranked in the upper half and living in the just world. In other words, they distort the probability of receiving positive feedback when being ranked in the upper half and living in the just world: $P(just) * \gamma_{RED} * P(F|\text{upper half}, just)$ with $\gamma_{RED} > 1$

Adapting the above derived Bayesian predictions (3.E.1) and (3.E.2) leads to the following deviations:[36]

$$Self\_Serving_{\text{IQ-test performance}} > Bayes\_IQ \tag{3.E.3}$$

$$Self\_Serving_{\text{unjust World}} < Bayes\_World \tag{3.E.4}$$

Condition B. BLUE types with negative Feedback:

BLUE types with negative feedback overstate the strength of the negative signal given that they are ranked in the upper half and living in the unjust world. In other words, they distort the probability of receiving negative feedback when being ranked in the upper half and living in the unjust world: $P(unjust) * \gamma_{BLUE} * P(F|\text{upper half}, unjust)$ with $\gamma_{BLUE} > 1$.

Adapting above derived Bayesian predictions (3.E.1) and (3.E.2), this leads to the following deviations: [37]

$$Self\_Serving\_IQ > Bayes\_IQ \tag{3.E.5}$$

$$Self\_Serving\_World > Bayes\_World \tag{3.E.6}$$

36. To see this, simply plug γ_{RED} into the two formulas. For the IQ-test performance belief both the numerator and the denominator are larger than in the Bayesian prediction. But as the numerator increases relatively more than the denominator the self-serving posterior is larger than the Bayesian prediction. For the unjust world belief (3.E.2) we have to plug in γ_{RED} only into the denominator. Then, the denominator is larger and the numerator is unchanged. As a consequence, self-serving RED types with positive feedback state lower beliefs.

37. To see this, simply plug γ_{BLUE} into the two formulas. For the IQ-test performance belief both the numerator and the denominator are larger than in the Bayesian prediction. But as the numerator increases relatively more than the denominator the self-serving posterior is larger than the Bayesian prediction. The same holds true for the unjust world belief (3.E.2). As the numerator increases relatively more than the denominator, self-serving BLUE types with negative feedback state a higher belief.

Hypotheses

Simply comparing the stated posteriors with the Bayesian prediction is insufficient. Numerous studies show that individuals often deviate from the Bayesian predictions even in the absence of motives. One commonly seen deviation is conservatism, i.e. individuals significantly understate the strength of received signals and thus update less than what Bayes' Rule would suggest. Take deviation (3.E.3) for example. If subjects in our treatment are both self-serving and conservative we might not detect a difference between the stated posterior and the Bayesian prediction and would erroneously reject the hypothesis that participants are self-serving. To rule out that observed effects are driven by other (non-self-serving) deviations from the Bayesian predictions, we introduced a control condition. We argue that subjects in our control group exhibit all the potential deviations except self-serving motives. Therefore, comparing the relative updating behavior between the subjects in the treatment and the control group provides casual evidence on self-serving attributions. To formalize this idea, we introduce yet another parameter δ, which incorporates all the deviations from the Bayesian predictions, i.e. conservatism, except the self-serving motive. We now can write down the stated posteriors for the control and the treatment group in the following way. In the control the stated beliefs are:

$$Posterior\_IQ = \delta * Bayes\_IQ$$
$$Posterior\_World = \delta * Bayes\_World$$

In the treatment the stated beliefs are:

$$Posterior\_IQ = \delta * Self\_Serving\_IQ$$
$$Posterior\_World = \delta * Self\_Serving\_World$$

Using the derived posterior beliefs in the treatment and control and the Bayesian predictions, we now derive the hypotheses for the two relevant cases.
Condition A. RED types with positive Feedback:

$$rel\_IQ_{Treatment} = \frac{\delta * Self\_Serving\_IQ}{Bayes\_IQ} > \frac{\delta * Bayes\_IQ}{Bayes\_IQ} = rel\_IQ_{Control} \quad (3.E.7)$$

In words, RED types that received positive feedback state a higher relative IQ-test performance Belief in the treatment than in the control group, i.e. they hold a more positive self assessment relative to the Bayesian prediction.. This holds true as (3.E.3) $\frac{Self\_Serving\_IQ}{Bayes\_IQ} > 1$.

$$rel\_World_{Treatment} = \frac{\delta * Self\_Serving\_World}{Bayes\_World} < \frac{\delta * Bayes\_World}{Bayes\_World} = rel\_World_{Control} \quad (3.E.8)$$

In words, RED types that received positive feedback state a lower relative unjust World Belief in the treatment than in the control group, i.e. they assume that it is less likely that they are in the unjust world relative to the Bayesian prediction. This holds true as (3.E.4) $\frac{Self\_Serving\_World}{Bayes\_World} < 1$.

Condition B. BLUE types with negative Feedback:

$$rel\_IQ_{Treatment} = \frac{\delta * Self\_Serving\_IQ}{Bayes\_IQ} > \frac{\delta * Bayes\_IQ}{Bayes\_IQ} = rel\_IQ_{Control} \quad (3.E.9)$$

In words, BLUE types that received negative feedback state a higher relative IQ-test performance Belief in the treatment than in the control group, i.e. they hold a more positive self assessment relative to the Bayesian prediction. This holds true as (3.E.5) $\frac{Self\_Serving\_IQ}{Bayes\_IQ} > 1$.

$$rel\_World_{Treatment} = \frac{\delta * Self\_Serving\_World}{Bayes\_World} > \frac{\delta * Bayes\_World}{Bayes\_World} = rel\_World_{Control} \quad (3.E.10)$$

In words, BLUE types that received negative feedback state a higher relative unjust World Belief in the treatment than in the control group, i.e. they assume that it is more likely that they are in the unjust world relative to the Bayesian prediction. This holds true as (3.E.6) $\frac{Self\_Serving\_World}{Bayes\_World} > 1$.

Appendix 3.F 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.[38] The next component is the real effort task, in which

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

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

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.

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:

---

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.

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)

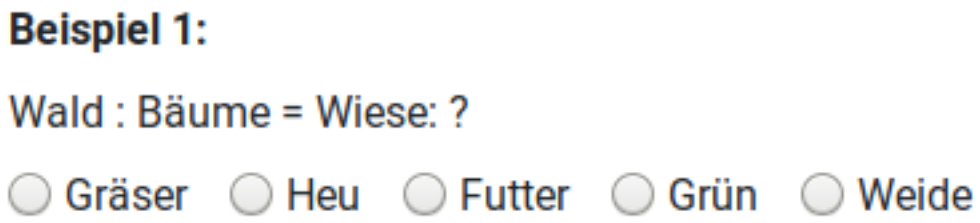
  - PhD
  - Different certificate
  - Prefer not to say

- 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

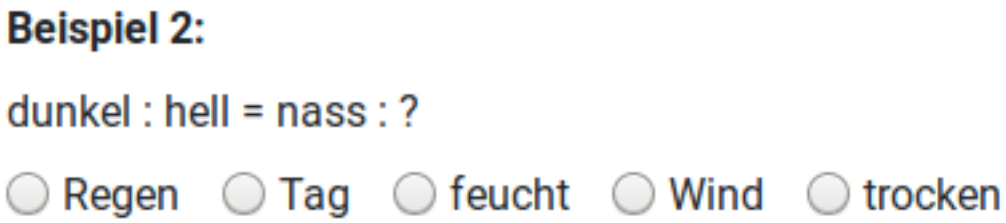
- What is your monthly household net-income?
- On a scale form 0 to 10, how willing are you to take risks? 0 means that you are not willing to take risks at all and 10 means that you are more than willing to do so.

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:



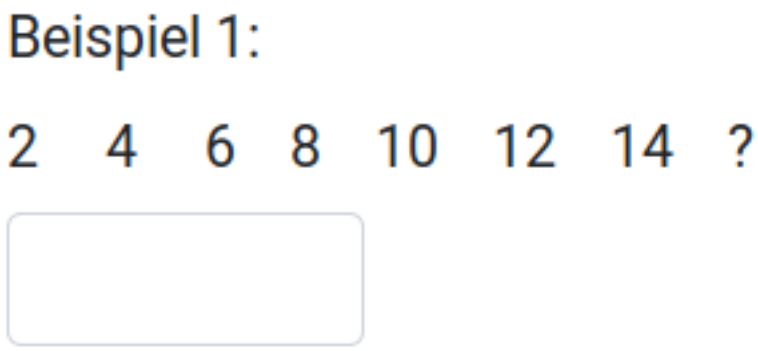
The relation between Wald (Forest) and Bäume (Trees) is that there are many trees in a forest. From the suggested options you now have to find the word that is similarly related to the third word Wiese (meadow). The correct answer is Gräser (Grass).



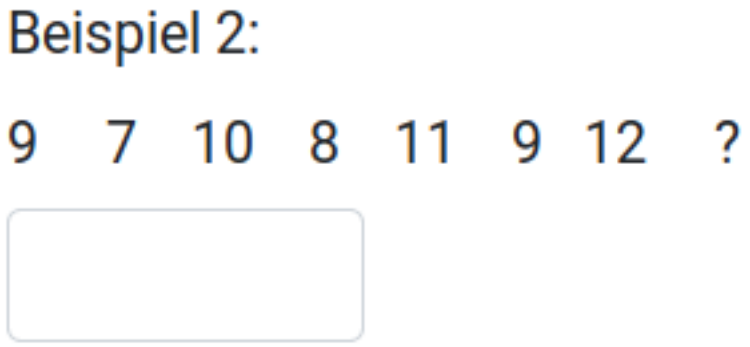
Dunkel (Dark) is the opposite of hell (bright), so you have to find the opposite of nass (wet). The right answer to example 2 is trocken (dry).

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: $9 \breve{} 2 = 7$; $7 + 3 = 10$; $10 \breve{} 2 = 8$; $8 + 3 = 11$; $11 - 2 = 9$; $9 + 3 = 12$; $12 - 2 = 10$. 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:

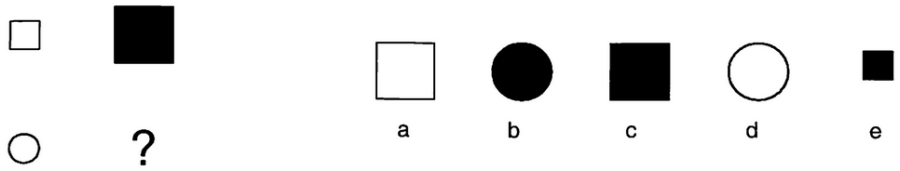
Your time is up. Part 2 of the test is over.

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:

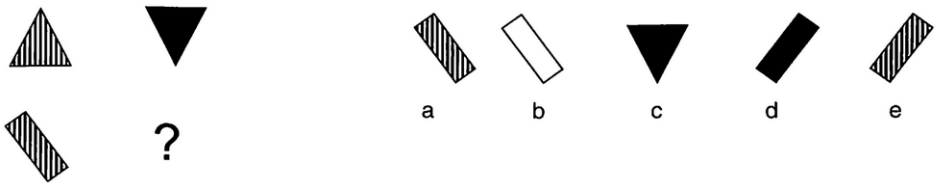
Beispiel 1:



○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

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. [178]

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. [120]

Barron, Kai. 2016. "Belief updating: Does the'good-news, bad-news' asymmetry extend to purely financial domains?" [120]

Benabou, Roland, and Jean Tirole. 2006. "Belief in a just world and redistributive politics." *Quarterly journal of economics* 121 (2): 699–746. [116]

Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." *Quarterly journal of economics* 117 (3): 871–915. [119, 120]

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. [120, 121]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg registration and organization online tool." *European Economic Review* 71: 117–120. [126]

Brunnermeier, Markus K, and Jonathan A Parker. 2005. "Optimal expectations." *American Economic Review* 95 (4): 1092–1118. [120]

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini. 2013. "Overconfidence and social signalling." *Review of Economic Studies* 80 (3): 949–983. [120]

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. [126]

Chen, Zhuoqiong Charlie, and Tobias Gesche. 2017. "Persistent bias in advice-giving." *University of Zurich, Department of Economics, Working Paper*, (228): [120]

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics* 22 (2): 369–395. [120, 129]

Coutts, Alexander, Leonie Gerhards, and Zahra Murad. 2019. "No one to blame: Self-attribution bias in updating with two-dimensional uncertainty." [121]

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. [120]

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. "Investor psychology and security market under-and overreactions." *Journal of Finance* 53 (6): 1839–1885. [120]

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying not to go to the gym." *American Economic Review* 96 (3): 694–719. [115]

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism." *American Economic Review* 105 (11): 3416–42. [120]

Dohmen, Thomas, and Armin Falk. 2011. "Performance pay and multidimensional sorting: Productivity, preferences, and gender." *American economic Review* 101 (2): 556–90. [115]

Doukas, John A, and Dimitris Petmezas. 2007. "Acquisitions, overconfident managers and self-attribution bias." *European Financial Management* 13 (3): 531–577. [120, 121]

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. [117, 118, 120, 122, 124, 136]

Epley, Nicholas, and Thomas Gilovich. 2016. "The mechanics of motivated reasoning." *Journal of Economic perspectives* 30 (3): 133–40. [119]

Exley, Christine L. 2016. "Excusing selfishness in charitable giving: The role of risk." *Review of Economic Studies* 83 (2): 587–628. [120]

Exley, Christine L. 2020. "Using charity performance metrics as an excuse not to give." *Management Science* 66 (2): 553–563. [120]

Exley, Christine L, and Judd B Kessler. 2019. "Motivated errors." Techreport. National Bureau of Economic Research. [120]

Falk, Armin, Thomas Neuber, and Nora Szech. 2020. "Diffusion of Being Pivotal and Immoral Outcomes." *Review of Economic Studies*, [120]

Gervais, Simon, and Terrance Odean. 2001. "Learning to be overconfident." *Review of Financial Studies* 14 (1): 1–27. [120, 121]

Gill, David, and Victoria Prowse. 2012. "A structural analysis of disappointment aversion in a real effort competition." *American Economic Review* 102 (1): 469–503. [117, 125]

Gneezy, Uri, Elizabeth A Keenan, and Ayelet Gneezy. 2014. "Avoiding overhead aversion in charity." *Science* 346 (6209): 632–635. [120]

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the self." *Games and Economic Behavior* 120: 311–324. [120]

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information avoidance." *Journal of Economic Literature* 55 (1): 96–135. [120]

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. [120]

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2018. "Unrealistic expectations and misguided learning." *Econometrica* 86 (4): 1159–1214. [119, 140]

Hestermann, Nina, and Yves Le Yaouanq. 2020. "Experimentation with Self-Serving Attribution Biases." *American Economic Journal: Microeconomics*, [119, 140]

Hilary, Gilles, and Lior Menzly. 2006. "Does past success lead analysts to become overconfident?" *Management science* 52 (4): 489–500. [120, 121]

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. [120, 121]

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. [120, 121]

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American economic review* 90 (4): 1072–1091. [120]

Köszegi, Botond. 2006. "Ego utility, overconfidence, and task choice." *Journal of the European Economic Association* 4 (4): 673–707. [119, 120]

Kőszegi, Botond, and Matthew Rabin. 2006. "A model of reference-dependent preferences." *Quarterly Journal of Economics* 121 (4): 1133–1165. [115]

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108 (3): 480. [119]

Lerner, Melvin J. 1980. "The belief in a just world." In *The Belief in a just World*. Springer, 9–30. [116]

Li, Feng. 2010. "Managers' self-serving attribution bias and corporate financial policies." *Available at SSRN 1639005*, [120, 121]

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. [120, 121]

Malmendier, Ulrike, and Geoffrey Tate. 2005. "Does overconfidence affect corporate investment? CEO overconfidence measures revisited." *European financial management* 11 (5): 649–659. [115]

Malmendier, Ulrike, and Timothy Taylor. 2015. "On the Verges of Overconfidence." *Journal of Economic Perspectives* 29 (4): 3–8. [115]

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, developmental, and cultural differences in the self-serving attributional bias." *Psychological bulletin* 130 (5): 711. [120]

Miller, Dale T, and Michael Ross. 1975. "Self-serving biases in the attribution of causality: Fact or fiction?" *Psychological bulletin* 82 (2): 213. [120]

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." Techreport. National Bureau of Economic Research. [118, 120]

Schneider, David J, Albert H Hastorf, and Phoebe Ellsworth. 1979. *Person perception*. Random House. [115]

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and self-deception." *Nature human behaviour* 3 (10): 1055–1061. [120]

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. [119, 120]

Stötzer, Lasse Simon. 2020. "Stereotypes about Refugees - how motives mold peoples' stereotypes." [120]

Topolewska-Siedzik, Ewa, Ewa Skimina, Włodzimierz Strus, Jan Cieciuch, and Tomasz Rowiński. 2014. "The short IPIP-BFM-20 questionnaire for measuring the big five." *Roczniki Psychologiczne // Annals of Psychology* 17: 385–402. [178]

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." *American Economic Review* 110 (2): 337–61. [120]

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. [120]