

Of Discussions, Beliefs, and Algorithms: Essays in Experimental Economics

Universität Hamburg

Fakultät für Wirtschafts- und Sozialwissenschaften

Kummulative Dissertation

Zur Erlangung der Würde des Doktors der
Wirtschafts- und Sozialwissenschaften

„Dr. rer. pol.“

(gemäß der PromO vom 18. Januar 2017)

vorgelegt von

Jan Biermann

aus St. Petersburg, Russland

Hamburg, den 22. Februar 2024

Vorsitzender: Prof. Dr. Andreas Lange

Erstgutachterin: Prof. Dr. Dr. Lydia Mechtenberg

Zweitgutachter: Prof. Dr. Robert Dur

Datum der Disputation: 12.09.2024

To the universal qualities of compassion and courage.

Acknowledgements

First, I would like to thank my supervisors. Thank you, Lydia (Mechtenberg), for your continued support and for believing in me. Thank you for engaging with my work and helping to improve it more than anyone else. Thank you, Robert (Dur), for all the motivating and inspiring interactions. Thank you for helping me to move forward many times when I was stuck with my projects.

Thank you, Johannes (Walter), for the amazing experience of working so closely together, thank you for teaching me more about economics than anyone else, and thank you for improvising with me. Thank you, Hendrik (Hüning), for being my big academic brother and for having my back in so many situations. Thank you, John (Horton), for providing me with many new perspectives and giving me the opportunity to visit MIT.

I am very aware of the enormously privileged employment situation during my PhD. These privileges include inspiring colleagues, various training programs and seminars, opportunities to attend conferences, and a research stay abroad as part of my PhD. Thanks to Lydia (Mechtenberg) and to everyone making the CDM-program possible, especially to Anke (Gerber). I would also like to thank thousands of people behind the data of my experiments who have enabled me to study human behavior.

I am deeply grateful to many people outside of academia who have greatly influenced me during the time of this dissertation. Thank you, Nicola, Sven, Anne, and Christos for having been my home. Thank you, Mom, for your love and every opportunity I have in life. Thanks to the improv community in Hamburg and ITM in Mannheim for teaching me how to play and so many other skills. A big thank you to all the people who have shown me profound ways to live a more examined and fulfilling life, especially to Sam, Adya, Sven, Nicolai, Laurenz, and Hadi.

Thank you!

Contents

Acknowledgements	iii
1 Introduction	1
2 Migrant Aid - an Experiment on the Effects of Online Debates	11
2.1 Introduction	12
2.2 Experimental Design	15
2.2.1 Chat vs. Notes	16
2.2.2 Trust Games	18
2.2.3 Attitudes and Beliefs	18
2.2.4 Pay-offs and Recruitment	19
2.3 Data	20
2.4 Results	22
2.4.1 <i>Chat</i> Stimulates Donations	22
2.4.2 A Priori Social Norm	23
2.4.3 Opinion-Expression Norm	23
2.4.4 Overly Optimistic Beliefs About Partner's Refugee Attitudes	27
2.4.5 Indirect Effect of <i>Chat</i> on Trust	29
2.5 Discussion	31
2.6 Conclusion	32
Bibliography	34
Appendix	38
3 Truth-telling and Wishful Thinking	48
3.1 Introduction	49
3.2 Hypotheses	52
3.3 Experimental Design	53

3.3.1	Experimental Task and Main Outcome Variable	53
3.3.2	Experimental Dimension I: HIGH vs. LOW Temptation to Lie . . .	54
3.3.3	Experimental Dimension II: ACTIVE vs. PASSIVE	56
3.3.4	Payment Scheme and Experimental Procedure	56
3.4	Results	57
3.5	Discussion	62
3.5.1	Alternative Mechanisms: False Consensus Effect	62
3.5.2	Alternative Mechanisms: Own Behavior vs. Behavior of a Random Participant	63
3.5.3	Caveats, Parallel Trend Assumption, and Lying as an Experience Good	64
3.5.4	Non-causal Evidence That Lying Causes More Negative Beliefs . . .	66
3.6	Conclusion	67
	Bibliography	74
	Appendix	75
4	Advised by an Algorithm: Assessing Advice Quality with Informa- tional Support and Reactions to Diverse Advice Quality	96
4.1	Introduction	98
4.2	Part I: Informational Support	103
4.2.1	Experimental Design	103
4.2.2	Data	107
4.2.3	Results	108
4.3	Part II: Diverse Advice Quality	110
4.3.1	Experimental Design	110
4.3.2	Results	111
4.4	Discussion	114
4.4.1	Informational Support	114
4.4.2	Diverse Advice Quality	118
4.4.3	Caveats and Potential Extensions	119

4.5 Conclusion	120
Bibliography	125
Appendix	126
Anhang der Dissertation	vii
Liste der aus dieser Dissertation hervorgegangenen Veröffentlichungen . . .	vii
Abstract	viii
Zusammenfassung	x
Erklärung	xii
Eidesstattliche Versicherung	xii
Selbstdeklaration bei kumulativen Dissertationen	xiii

Chapter 1

Introduction

Although research in economics was once regarded as typically non-experimental,¹ current modern economic research frequently includes experiments as a well-established method (Reuben et al., 2022). As with any other method, experimental research has advantages as well as disadvantages, the most apparent advantage being a controlled environment, which allows for strong causality claims, and the most prominent disadvantage being a lack of external validity, which makes generalizations challenging. This dissertation employs experiments and leverages the strengths of such methods while maintaining critical awareness of their limitations. Through the lens of experimental methods, it provides novel insights into (behavioral) economic concepts of trust, generosity, belief formation, and advice taking. The underlying view of economics as a discipline in this dissertation stands in the tradition of Richard Thaler who fundamentally shaped the modern understanding of behavioral economics by emphasizing the integration of psychological concepts and stressing the importance of evidence-based methods (Thaler, 2016).

The essays in this dissertation are connected in three different ways. First, all three essays use experimental methods. Chapter 2 reports on a lab-in-the-field experiment (Gneezy and Imas, 2017) studying the behavior of German adolescents. Chapters 3 and 4 relate to the use of online experiments with participants from the United Kingdom (chapter 3) and the United States (chapter 4). Second, beliefs play a key role throughout the dissertation. The formation of beliefs about the behavior of others is the central topic of chapter 3.

¹As recently as 1985 Samuelson and Nordhaus wrote that “economists cannot perform the controlled experiments of chemists or biologists” (Samuelson and Nordhaus, 1985, p. 8).

In chapter 2 and 4, we analyze how much an individual trusts a person with whom they have just discussed a potentially controversial topic (chapter 2) and how much a person is willing to rely on advice provided by an algorithm (chapter 3). In both cases, underlying beliefs (about the other person or the algorithm) are likely to play a key role. Third, the dissertation gives different angles on collective decision-making in each of the three essays. Chapter 2 examines the difference between making a decision alone and discussing it with a peer. The setting in chapter 4 can be interpreted as a collective decision-making situation involving a human and an automated agent, i.e., an algorithm, in which the final authority to decide resides with the human. Chapter 3 implicitly touches on this topic by addressing the formation of beliefs about others which can be considered the basis for communication and cooperation. While all essays primarily address academic economists, they also speak to scholars in other disciplines, most importantly in political science (chapter 2), computer science (chapter 4), and social psychology (chapters 2, 3, 4).

Chapter 2 (co-authored with Hendrik Hüning and Lydia Mechtenberg) examines a debating process that precedes a collective decision. It is evident that the statements made in such debates are at least partly determined by the desired outcomes. However, the debaters' suggestions might also be motivated by reputational concerns. The arguments they articulate signal their personal attitudes, preferences, and beliefs. Further, people exposed to arguments in favor of, say, pro-social behavior are more likely to signal a high willingness to behave altruistically as well, thereby signaling conformity to the preferences of the others. Subsequently, such dynamics can shape the collective decision, e.g., through individuals' commitment to their own expressed opinions. Our experiment examines the role of such pre-vote deliberation in situations where members of a privileged group of German school minors decide how much of their resources they will transfer to a disadvantaged group, i.e., refugee minors. We focus on two related research questions: First, how does the discussion on sharing funds with same-age refugees affect generosity, i.e., the size of transfers? Second, how does this discussion affect trust between the discussants?

We conduct a lab-in-the-field experiment with 488 school minors from 13 schools in Germany, 96% of whom are natives. We randomly assign school minors into pairs to discuss via chat how much of a windfall gain that would be part of their class fund to donate to a charity supporting incoming refugee minors. Before and after the discussion and collective decision on migrant aid the matched pairs play a trust game (Berg, Dickhaut, and McCabe, 1995). In the control treatment, participants privately state and give reasons for their individual preference regarding a decision on migrant aid, while the decision itself remains collective. Our main finding is that the chat discussion about transferring resources to refugees increases the size of the migrant aid by more than 14%. Further, the chat discussion contains less negative sentiment and fewer expressed negative attitudes toward refugees compared to the benchmark treatment with individual reasoning, i.e., some negative attitudes are suppressed. Additionally, we find that the chat discussion does not affect trust directly; but through changes in beliefs about the co-players' attitudes toward refugees. In other words, school minors who are perceived as refugee-friendly after the discussion win increased trust from their chat partner.

The essay speaks to several strands of literature. First, it is related to the literature covering social information on donation behavior (e.g., Frey and Meier, 2004; Martin and Randal, 2008; Croson and Shang, 2008; Shang and Croson, 2009). Second, it is connected to studies that treat generosity as a signal of trustworthiness (e.g., Gambetta and Przepiorka, 2014; Fehrler and Przepiorka, 2016). Third, it relates to contributions from other disciplines on how expressing altruism can boost a person's reputation (e.g., Wedekind and Braithwaite, 2002; Bereczkei, Birkas, and Kerekes, 2010; Mocos and Scheuring, 2019). None of the existing studies features an interactive chat as implemented in our work. This innovation constitutes a richer setting than one finds in restricted communication and allows us to consider different effects of direct social interaction between the decision-makers in our study. This paper also contributes to the literature on economic experiments with children and adolescents (for an overview, cf. Sutter, Zoller, and Glätzle-Rützler, 2019).

Chapter 3 focuses on the formation of beliefs about the behavior of others. It acknowledges that beliefs are not always rational, but can be “motivated” in that they serve a

psychological function of maintaining a positive self-perception. In this sense, an individual's beliefs about others can be driven by their own (past) behavior. The perceptions one holds about others are fundamental, as they form the foundation for potential engagement and cooperation with fellow humans. Specifically, this paper is concerned with whether breaking a norm negatively affects what people think others would do in the same situation. In this study, the relevant social norm is that of acting honestly (Bicchieri, 2005).

I use an online experiment with a representative sample of the UK's general population (1201 participants). Following a novel experimental design, subjects are exposed to one of two conditions. Lying is possible in both cases, but it is more tempting in one situation than in the other. This study finds that if lying is more tempting, the proportion of people who choose to lie increases. Importantly, subjects in this condition also hold more pessimistic beliefs about other people. I include an additional treatment dimension, in which subjects are not actively involved in the situation, i.e., they have no opportunity to lie. Including the beliefs of these passive subjects in the analysis enables me to argue that the differences in beliefs are not due to a rational anticipation of the incentive structure.

By establishing a novel way in which humans seek to rationalize past actions, this paper contributes to the literature on motivated information processing and strategic belief manipulation (e.g., Ehrich and Irwin, 2005; Dana, Weber, and Kuang, 2007; Di Tella et al., 2015; Ging-Jehli, Schneider, and Weber, 2020; Bicchieri, Dimant, and Sonderegger, 2023; Mechtenberg et al., 2024). In my setting, participants justify past actions by adjusting their beliefs about whether others conform to the relevant social norm in the same situation. The paper addresses the social norm of acting honestly and, hence, speaks to the sizable literature on lying (for an overview, cf. Abeler, Nosenzo, and Raymond, 2019).

Chapter 4 (co-authored with John J. Horton and Johannes Walter) examines a situation in which a human decision-maker is assisted by an algorithm. Such settings are becoming increasingly common as is evident in areas as diverse as healthcare, the judiciary, and e-commerce, where, respectively, physicians use algorithms to determine the most suitable treatments for patients, judges rely on them to support sentencing decisions, and

pricing managers utilize them to strategically set discounts for products. While existing literature has established that humans react differently to human vs. algorithmic advice (e.g., Dietvorst, Simmons, and Massey, 2015; Logg, Minson, and Moore, 2019; Prahla and Van Swol, 2021; Sele and Chugunova, 2024), more research is required to better understand the mechanics of human responses to algorithmic recommendations. Our work contributes to this endeavor by focusing on two aspects: First, the study examines informational resources that are designed to aid individuals in evaluating algorithmic guidance, and second, it explores human reactions to varying algorithmic performance.

To address these questions, we conduct an online experiment involving 1565 participants from the US. In the baseline treatment, subjects repeatedly perform an estimation task and are provided with algorithmic guidance, all without them knowing what type of algorithm is used and without receiving feedback. Our treatments introduce two interventions aimed at enhancing the quality of human decisions when receiving algorithmic advice. In the first intervention, we explain how the algorithm functions. We find that while this intervention reduces adherence to the algorithmic advice, it does not improve decision-making performance. In the second treatment, we disclose the correct answer to the task after each round. This intervention reduces adherence to algorithmic advice and improves human decision-making performance. Further, we investigate the extent to which individuals are capable of adjusting their assessment of an algorithm when the advice quality varies due to external circumstances. We find some evidence that people adjust their adherence depending on the quality of the algorithmic recommendations.

We contribute to the existing literature in two major ways. First, we test interventions designed to improve people’s assessment of algorithmic advice. Importantly, our interventions are informed by the existing literature on decision-making under risk. We introduce the concept of learning through thought vs. experience to the literature on empirical human-AI decision-making. In doing so, we bring more nuance to the debate on how people learn about algorithms. This contribution speaks to the literature which empirically explores how to calibrate trust in an algorithm (e.g., Yin et al., 2019; Zhang, Liao, and Bellamy, 2020; Alufaisan et al., 2021; Park et al., 2019; Green and Chen, 2019).

Second, we investigate the extent to which individuals can skillfully adjust their evaluations of the algorithm in response to varying levels of advice quality. In this regard, we contribute by showing that subjects do not necessarily abandon the algorithm after having seen it make poor predictions. Rather, if the low-quality advice is caused by the algorithm being inappropriate for a certain setting, humans will continue to use it under different circumstances. These results contribute to the literature on how humans react to algorithmic mistakes, including the seminal paper by Dietvorst, Simmons, and Massey (2015) and subsequent related studies (e.g., Prahla and Van Swol, 2017; Dietvorst, Simmons, and Massey, 2018; Zhang and Gosline, 2022).

Jointly, the three experimental essays shed light on the different roles that fellow humans and non-human agents play in economic decision-making. The combined work also provides new insights relevant to the understanding of collective decision-making. Chapter 2 examines the effects of debates among peers on generosity and trust. Chapter 3 shows how norm-breaking can influence an individual’s view of others. Chapter 4 analyzes the factors that contribute to the acceptance of algorithmic advice.

Bibliography

- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. “Preferences for truth-telling.” *Econometrica* 87 (4):1115–1153.
- Alufaisan, Yasmineen, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. “Does explainable artificial intelligence improve human decision-making?” *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (8):6618–6626.
- Berczkei, Tamas, Bela Birkas, and Zsuzsanna Kerekes. 2010. “Altruism towards strangers in need: Costly signaling in an industrial society.” *Evolution and Human Behavior* 31:95–103.

- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. “Trust, reciprocity, and social history.” *Games and Economic Behavior* 10 (1):122–142.
- Bicchieri, Cristina. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.
- Bicchieri, Cristina, Eugen Dimant, and Silvia Sonderegger. 2023. “It’s not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion.” *Games and Economic Behavior* 138:321–354.
- Croson, Rachel and Jen Shang. 2008. “The impact of downward social information on contribution decisions.” *Experimental Economics* 11 (3):221–233.
- Dana, Jason, Roberto A. Weber, and Jason X. Kuang. 2007. “Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness.” *Economic Theory* 33:67–80.
- 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):3416–3442.
- Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. “Algorithm aversion: people erroneously avoid algorithms after seeing them err.” *Journal of Experimental Psychology: General* 144 (1):114.
- . 2018. “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them.” *Management Science* 64 (3):1155–1170.
- Ehrich, Kristine R. and Julie R. Irwin. 2005. “Willful ignorance in the request for product attribute information.” *Journal of Marketing Research* 42 (3):266–277.
- Fehrler, Sebastian and Wojtek Przepiorka. 2016. “Choosing a partner for social exchange: Charitable giving as a signal of trustworthiness.” *Journal of Economic Behavior and Organization* 129:157–171.

- Frey, Bruno S. and Stephan Meier. 2004. “Social comparisons and pro-social behavior: Testing ”conditional cooperation” in a field experiment.” *American Economic Review* 94 (5):1717–1722.
- Gambetta, Diego and Wojtek Przepiorka. 2014. “Natural and strategic generosity as signals of trustworthiness.” *PloS one* 9 (5):e97533.
- Ging-Jehli, Nadja R., Florian H. Schneider, and Roberto A. Weber. 2020. “On self-serving strategic beliefs.” *Games and Economic Behavior* 122:341–353.
- Gneezy, Uri and Alex Imas. 2017. “Lab in the field: Measuring preferences in the wild.” In *Handbook of Economic Field Experiments*. 439–464.
- Green, Ben and Yiling Chen. 2019. “The principles and limits of algorithm-in-the-loop decision making.” *Proceedings of the ACM on Human-Computer Interaction* 3:1–24.
- Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. “Algorithm appreciation: People prefer algorithmic to human judgment.” *Organizational Behavior and Human Decision Processes* 151:90–103.
- Martin, Richard and John Randal. 2008. “How is donation behaviour affected by the donations of others?” *Journal of Economic Behavior & Organization* 67 (1):228–238.
- Mechtenberg, Lydia, Grischa Perino, Nicolas Treich, Jean-Robert Tyran, and Stephanie W. Wang. 2024. “Self-signaling in voting.” *Journal of Public Economics* Forthcoming.
- Mokos, Judit and István Scheuring. 2019. “Altruism, costly signaling, and withholding information in a sport charity campaign.” *Evolution, Mind and Behaviour* 17 (1):10–18.
- Park, Joon S., Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. “A slow algorithm improves users’ assessments of the algorithm’s accuracy.” *Proceedings of the ACM on Human-Computer Interaction* 3:1–15.
- Prahl, Andrew and Lyn M. Van Swol. 2017. “Understanding algorithm aversion: When is advice from automation discounted?” *Journal of Forecasting* 36 (6):691–702.

- . 2021. “Out with the humans, in with the machines? Investigating the behavioral and psychological effects of replacing human advisors with a machine.” *Human-Machine Communication* 2:209–234.
- Reuben, Ernesto, Sherry X. Li, Sigrid Suetens, Andrej Svorenčík, Theodore Turocy, and Vasileios Kotsidis. 2022. “Trends in the publication of experimental economics articles.” *Journal of the Economic Science Association* 8 (1):1–15.
- Samuelson, Paul A. and William D. Nordhaus. 1985. *Economics*. McGraw-Hill.
- Sele, Daniela and Marina Chugunova. 2024. “Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making.” *Plos one* 19 (2):e0298037.
- Shang, Jen and Rachel Croson. 2009. “A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods.” *The Economic Journal* 119 (540):1422–1439.
- Sutter, Matthias, Claudia Zoller, and Daniela Glätzle-Rützler. 2019. “Economic behavior of children and adolescents – A first survey of experimental economics results.” *European Economic Review* 111:98–121.
- Thaler, Richard H. 2016. “Behavioral economics: Past, present, and future.” *American Economic Review* 106 (7):1577–1600.
- Wedekind, Claus and Victoria A. Braithwaite. 2002. “The long-term benefits of human generosity in indirect reciprocity.” *Current Biology* 12:1012–1015.
- Yin, Ming, Vaughan Wortman, Jennifer Wortman, and Hanna Wallach. 2019. “Understanding the effect of accuracy on trust in machine learning models.” *Proceedings of CHI Conference on Human Factors in Computing Systems* :1–12.
- Zhang, Yunfeng, Vera Liao, and Rachel K. E. Bellamy. 2020. “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

Zhang, Yunhao and Renee Gosline. 2022. “Understanding Algorithm Aversion: When Do People Abandon AI After Seeing It Err?” *SSRN Working Paper* .

Chapter 2

Migrant Aid - an Experiment on the Effects of Online Debates

Authors: Jan Biermann, Hendrik Hüning and Lydia Mechtenberg

Abstract

In a lab-in-the-field experiment with 488 school minors from 13 schools, we randomly assign school minors into pairs that discuss and decide how much of a common fund to transfer to refugee minors. The discussion takes place in a digital chat environment. A treatment without discussion serves as a benchmark. Before and after, pairs play a trust game that allows us to measure changes in trust due to the discussion and collective decision on the (experimental) migrant aid. We find that discussing the size of this migrant aid increases transfers by more than 14%. Moreover, school minors that are perceived as refugee-friendly after the discussion benefit from increased trust by their chat-partner. Our results hold relevance for how social information affects migrant aid and how the expressed willingness to support transfers to migrants serves as a signal of trustworthiness.

Keywords: Communication, collective decisions, migrant aid, trust, text data, generosity

JEL: C93, D70, D83

2.1 Introduction

Rich countries often find themselves in a position where they must decide on transfers to a disadvantaged external group or country. Examples include the global distribution of COVID-19 vaccines (COVAX Initiative), resources for disaster relief, and transfers to incoming migrants and refugees. In all of these examples, the final decision is not simply an aggregation of individual preferences; rather, the debating process preceding the decision-making plays a vital role. Within the last decade, social media have been shaping such debates increasingly. Especially the younger generation engages in social and political issues more often online than offline (Nelson, Lewis, and Lei 2017). Hence, an increasingly large share of the population takes an active part in debates vital for collective decision-making on how much of its resources a society is willing to transfer to the disadvantaged that do not, or not yet, belong to it.

It is evident that the statements made in such debates are at least partly defined by the desired outcomes, such as determining the extent of resource sharing. However, it is important to note that the debaters' suggestions may also be motivated by reputational concerns. The arguments they communicate serve as signals of their personal attitudes, preferences, and beliefs. For instance, being exposed to arguments in favor of sharing resources graciously may make a person more likely to signal a high willingness to support transfers, too, in order to signal conformity to the political and social preferences of others. Importantly, such dynamics can shape the subsequent political support of transfers, e.g., due to a commitment to one's own expressed opinions. Our paper examines the role of such pre-vote deliberation in situations where members of a privileged group, i.e., German school minors, decide how much of their resources to transfer to a disadvantaged group, i.e., refugee minors.

In a lab-in-the-field experiment with 488 school minors from 13 schools in Germany, of whom 96% are natives, we study two related questions: First, how does the discussion on sharing funds with same-age refugees affect the size of transfers? Second, does this

discussion on migrant aid affect trust between the discussants?

In our experiment, we randomly assign school minors into pairs of two to discuss via chat how much of a windfall gain that would be part of their class fund to donate to a charity supporting incoming refugee minors. Before and after the discussion and collective decision on migrant aid the matched pairs play a trust game (Berg, Dickhaut, and McCabe 1995). This allows us to measure how trust changes due to the discussion and decision about the migrant aid. In a benchmark treatment without the chat, participants state their individual preference for a decision on migrant aid and reasons thereof privately, while the decision itself remains collective. Mimicking the societal decision problem of how much of their *common* resources to transfer, our subjects do not receive their earnings as individual payouts; instead, the earnings are disbursed into a class fund of the respective student. Importantly, since participants learn about the three components - trust game, chat and vote on migrant aid, and another trust game - one at a time, they cannot act strategically in one game to (potentially) benefit in a subsequent game.

Our main finding is that the chat discussion about transferring resources to refugees increases the size of the migrant aid by more than 14%. Additionally, the chat discussion contains less negative sentiment and fewer negative expressed attitudes towards refugees compared to the benchmark treatment with individual reasoning, i.e., some negative attitudes are suppressed. Moreover, we find that the chat discussion does not directly affect trust but through changes in beliefs about the co-player's attitudes towards refugees. In other words: The more a participant updates their belief about their co-player's attitudes towards refugees, believing them to be more refugee-friendly than initially suspected, the more trust increases.¹

To our knowledge, this is the first paper that experimentally investigates the role of online debates in shaping political support of migrant aid and trust among the debaters. Our results on how online debates influence migrant aid contribute to a well-established literature on the effects of social information on donation behavior (e.g., Frey and Meier

¹This increase in trust is rational since, in our sample, positive attitudes towards refugees are positively correlated with trustworthiness.

2004, Martin and Randal 2008, Croson and Shang 2008, Shang and Croson 2009). While most previous studies exogenously provide information about donations of others, the chat implemented in our experiment allows us to study a richer setting in which social information plays out to influence one's decision on which size of migrant aid to support. For instance, we find that the average first transfer suggestion in the chat is not statistically different from the average transfer of those participants who did not chat. Hence, the final decision to transfer more than first suggested must be a result of a deliberation process during the chat interaction rather than the fact of being matched with another person.

Second, other than in the literature on donations, the decision on how much of a common fund to share with a disadvantaged group is a collective decision in our setting. The paper therefore bridges the gap between the literature on charitable giving and the literature on collective decision-making and behavior toward other groups (e.g., Kranton et al. 2020).

Third, our results contribute to the literature on generosity and charitable giving as signals of trustworthiness, cf. Gambetta and Przepiorka (2014) and Fehrler and Przepiorka (2016). While the former study investigates the effect of natural versus strategic generosity towards peers on trust among them, the latter studies generosity towards a charity and its effect on trust among peers. Neither of these contributions, however, implements a chat that allows investigating the social interaction itself that leads to those effects. The implementation of the chat makes our experiment particularly suited to investigate the effects of value-laden online debates.

Finally, our paper contributes more generally to the literature on the signaling effect of altruism. Expressing altruism has been found to signal an individual's willingness to cooperate and can boost their reputation, e.g., within a community (Mokos and Scheuring 2019, Bhogal 2021). Further, literature has shown that it is not only altruism directed at peers that increases one's reputation, but also charitable giving (Bereczkei, Birkas, and Kerekes 2010, Wedekind and Braithwaite 2002). This effect can be even stronger than donations to peers (Milinski, Semmann, and Krambeck 2002). Therefore, the discussion

on migrant aid in our experiment can be interpreted as an exercise in enhancing one's reputation. Those who express a greater willingness to support larger transfers benefit in the subsequent trust game because they are perceived as more trustworthy. This is even more telling as, in our setting, this cannot be the result of strategic behavior.

The remainder of this paper is structured as follows: Section 2 introduces our experimental design and procedures. While section 3 presents the data, our empirical results are summarized in section 4. Section 5 provides a discussion of our findings in the light of related literature as well as the study's limitations. Section 6 concludes.

2.2 Experimental Design

We conducted our experiment with school minors during regular course hours in classrooms or computer rooms at our subjects' schools. Since the experiment contained activities such as online chatting and donating, which are part of real life, we considered a natural environment appropriate.

The experiment was entirely computer-based. The program was designed using the software oTree (Chen, Schonger, and Wickens 2016). At the beginning of each experimental session, school minors were randomly matched into pairs of two, which remained fixed throughout the experiment. We matched them across different classes or schools whenever we could allocate the same time slot to two different classes or even schools (see Table 2.1 for details). Moreover, we told all our subjects that they might have been matched with a school minor from a different class or even from a different school. Hence, we implemented a stranger matching among peers.

The six stages of the experiment are illustrated in Figure 2.1: First, in stage 1, the participants fill out a survey and watch one of three randomized videos intended to prime them in their attitude toward refugee minors. Then, in stage 2, they play a trust game with their matched co-player, to elicit basic trust toward the peer group. Afterwards, we treat our participants: They either communicate with their co-player or take notes (stage

3), both times on the pros and cons of donating to refugee minors. Next, in stage 4, they take the collective donation decision, each of them stating their preferred size of the migrant aid, followed by the computer randomly selecting one of the two. Finally, they play the trust game again (stage 5). Thereby, we elicit how trust may have changed due to what they have learned in the chat about the co-player's refugee-friendliness. Finally, they fill out a questionnaire. At the end of the experiment, we debrief our participants. The debriefing explicitly addresses the video prime.

Our design allows us to study the young generation dealing with the collective decision on how much of a common fund to share with a disadvantaged out-group. In particular, it allows us to test how in-group deliberation affects both this decision and changes in in-group trust.

2.2.1 Chat vs. Notes

In the two key stages (3 and 4) of our experiment, subject pairs are confronted with the collective choice of a donation to refugee minors. Each pair receives a joint experimental budget of 30 ECU (experimental currency units, called "Taler" in our experiment) as a windfall gain. Next, subjects are asked: *How much, if anything, of your common budget do you want to donate to help minor unaccompanied refugees?* Our participants have the option to click on a link to get more information about the donation. The money is donated to the project "Helping Refugee Children" by the organization "Deutsches Kinderhilfswerk" which is a well-known charitable organization in Germany.

Our subjects prepare for this decision by either chatting with each other (in the chat treatment) or by taking notes privately (in the notes treatment). In the former (*Chat*), subjects discuss the donation decision with their random partner in an online environment similar to WhatsApp; in the latter treatment (*Notes*), subjects take private notes in a similarly designed textbox. Both the chat and the note-taking last for seven minutes. Afterwards, both members of the matched pair enter their preferred donation into an entry field. This is both costless and mandatory.

The decision is implemented according to random dictatorship (Gibbard 1977): The preferred choice of one of the two co-players is randomly chosen and each choice has equal probability. The selected amount is donated to refugee minors. The remaining share of the joint experimental budget is equally split among the two co-players. By employing random dictatorship, we ensure that the subjects face incentives to reveal their true preferences rather than choosing strategically. The co-player's preferred choice and the final choice selection are revealed only at the very end of the experiment.

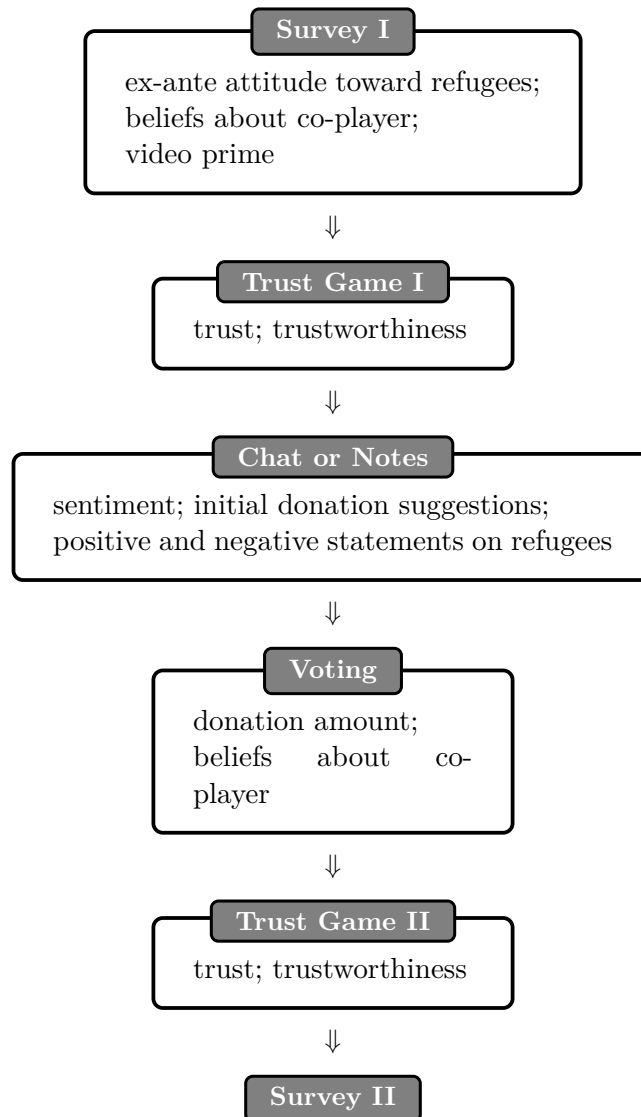


Figure 2.1 Stages of experiment including key variables

2.2.2 Trust Games

Apart from testing the effect of deliberation on collective sharing decisions, we investigate its potential effects on trust among the deliberating partners. For this purpose, we implement one trust game directly before and another trust game directly after the chat, respectively the notes stage. These trust games are in the spirit of Berg, Dickhaut, and McCabe (1995): One of the two players is randomly assigned to receive an endowment of 10 ECU and chooses how much, if anything, of this amount to send to the other player. The amount sent is tripled. The receiving player on their part decides how much, if anything, of the amount received to send back. Every subject makes a decision as a trustor (first mover) and ten decisions as a trustee (strategy method for the second mover, one decision for every possible decision of the trustor) in every trust game stage. Hence, we collect information on how much the trustee wants to send back given any hypothetical amount the trustor can send (trustworthiness). Every subject is paid based either on their decisions as a trustor or based on their decisions as a trustee, which is drawn randomly. The results of all trust games are revealed only at the very end of the experiment. Subjects do not know beforehand that the experiment contains a second trust game. Hence, we can rule out that their behavior during the first trust game and in the chat is influenced by the anticipation of the second trust game.

2.2.3 Attitudes and Beliefs

We elicit control variables in surveys directly before and after the experiment. Besides standard demographic questions, we ask participants how pronounced their positive and negative attitudes towards refugees are on a 5-point Likert scale ranging from *very strongly* to *weakly or not at all pronounced*. Moreover, we ask participants how pronounced they think the positive (negative) attitudes towards refugees of their co-player are. We elicit these beliefs at the beginning and the end of our study and are therefore able to measure the change in beliefs. Beliefs are incentivized such that a subject is rewarded a bonus of

0.5 ECU if they guessed correctly (see appendix for details). Beliefs about the refugee-friendliness of the co-player are important as potential mediators of trust between peers.

Furthermore, each participant is randomly assigned to one of three different videos on refugee minors at school (positive, negative, or balanced) during the first survey. The purpose of including these videos is to increase the heterogeneity of opinions as well as provide diverse arguments to our subjects that they could use in the chat or in their notes, depending on the treatment. The video primes have no significant effects, however.

2.2.4 Pay-offs and Recruitment

Pay-offs are determined with equal probability from the first trust game, the voting stage, or the second trust game. If the voting stage is drawn, the amount that remains after the donation is split equally between the two members of the respective pair. This pay-off structure is explained to the subjects at the beginning of the experiment, while the drawing takes place at the very end of the experiment. Participants are paid 0.6 Euro for each ECU they have earned in the experiment. Importantly, a participant's pay-off at the end of the experiment is not paid to the participants individually, but into the class fund. The class fund is a public good available to the class as a whole. Such resources are typically used to finance cultural activities of the class. Hence, in the voting stage, the participants make a decision between allocating resources to group insiders (class fund) or to group outsiders (refugee charity).

For the experiment, we contacted 214 secondary schools from the states of Saxony, Berlin, and Hamburg. In total, 16 schools agreed to collaborate with us, all located in Berlin or Hamburg. Hence, we have a selected sample of schools open to collaboration with researchers from Hamburg and to the topics of deliberation and refugees. School minors with averse views on refugees are most likely under-sampled. We conducted our experiment between December 2019 and March 2020 in thirteen of these schools. Four of those are located in Berlin and nine are located in Hamburg. In March, we had to stop our

fieldwork because of school closings due to the COVID-19 pandemic. For this reason, we could not conduct our experiment in the remaining three schools. We only considered a class if all pupils in it were at least fifteen years old; and we only accepted school minors whose parents gave informed consent, in addition to theirs.

2.3 Data

Overall, 501 school minors participated in our experiment in 19 different sessions. As our power calculation in the appendix reveals, this was the targeted quantity for our research study. Due to technical malfunction, we had to dismiss data from 13 subjects, resulting in 488 observations for our analysis. Table 2.1 presents some descriptive characteristics of our sample. Subjects are between 15 and 21 years old, averaging 17 years; 56% of subjects are female and one subject is diverse. From all participating school minors, 27% went to schools in Berlin. Almost all subjects are born in Germany (96%). On average, the trustors sent 5.52 Taler in trust game 1. On average, for each Taler their trustor sent, trustees returned 1.25 Taler in trust game 1. Our measure for subjects' general willingness to donate is their individual donation to Médecins sans Frontières (MSF), which we elicited in stage 1. It averages 10.11 Taler (6.06 Euro). An (incentivized) question asked participants how much of 10 Euro they would donate to Médecins Sans Frontières (MSF) if they were drawn as a winner of a lottery at the end of the session. The lottery randomly picked one school minor in each session to win 10 additional Euro on top of the final payoff. The maximum possible donation amount to refugee minors is 30 Taler and the average donation is 20.28 Taler (see Figure 2.9 for the distribution per treatment). Two-thirds of school minors are randomly assigned to the chat treatment and one-third to the notes treatment. Table 2.1 also reveals that positive attitudes are more pronounced than negative attitudes towards refugees. This allows for considerable heterogeneity between partners in *how* positive they feel toward refugees. Negative attitudes are less prevalent, which allows for less heterogeneity in the intensity of negative attitudes. These ex-ante attitudes are not different across treatments (MWU-test, p-values: 0.188

and 0.663, respectively).

The two minors in each matched pair were either from two different schools (29% of all pairs), from two different classes (28%), or from the same class (43%). However, our instructions did not explicitly mention the three possibilities; we only informed our participants that the partner could be from a different school. Our impression from the interaction with the minors during the debriefing is that most participants believed to have played the game with a partner from a different school.

Table 2.1 Summary statistics

Variable	Overall Mean	Notes Mean	Chat Mean	St. Dev.	Min	Max
Age	17	16.99	17	0.91	15	21
Share female	0.56	0.56	0.56	0.50	0	1
Share school in Berlin	0.27	0.28	0.27	0.44	0	1
Share born in Germany	0.96	0.97	0.95	0.20	0	1
Household members	2.75	2.75	2.74	1.11	0	7
Pocket money (in Euro)	24.89	29.53	22.41	39.36	0	450
Share same class	0.43	0.39	0.46	0.50	0	1
Share same school different class	0.28	0.28	0.28	0.45	0	1
Share different school	0.29	0.33	0.27	0.45	0	1
Positive attitudes refugees	3.60	3.68	3.54	0.94	1	5
Negative attitudes refugees	2.14	2.16	2.13	0.96	1	5
Taler sent trust game 1	5.52	5.55	5.49	2.66	0	10
Taler sent trust game 2	5.75	5.83	5.71	2.93	0	10
Amount returned trust game 1	1.26	1.24	1.27	0.42	0	3
Amount returned trust game 2	1.25	1.23	1.27	0.43	0	3
Donation to MSF	10.11	10.52	9.86	6.67	0	16.67
Donation to refugees	20.28	18.64	21.19	8.86	0	30
Share positive video	0.34	0.32	0.35	0.47	0	1
Share negative video	0.33	0.32	0.32	0.47	0	1
Share balanced video	0.34	0.36	0.33	0.47	0	1
Share chat treatment	0.64	0	1	0.48	0	1
Pay-off (in Euro)	5.73	6.07	5.54	3.49	0	19

Notes: The number of observations is 488. The variables “Amount returned” correspond to the amount of Taler sent back by the trustee for each Taler received. Reported amounts are in Taler unless stated otherwise.

2.4 Results

2.4.1 *Chat* Stimulates Donations

Our analysis reveals that participants in *Chat* show a significantly higher willingness to donate to refugees, compared to participants in *Notes*. Results are summarized in the right panel of Figure 2.2. While those in *Chat* donate on average 21.22 Euro, those in *Notes* donate 18.59 Euro, a difference of more than 14% (MWU-test, p-values: 0.023). As a robustness check, we also perform Tobit regressions with individual donations as the dependent variable that find the same results, see appendix. We correct the p-values of our direct treatment effects, namely the effect of *Chat* on donations, beliefs, and trust as well as the effect of the video primes on donations for multiple hypotheses testing using the Holm-method (Holm 1979).

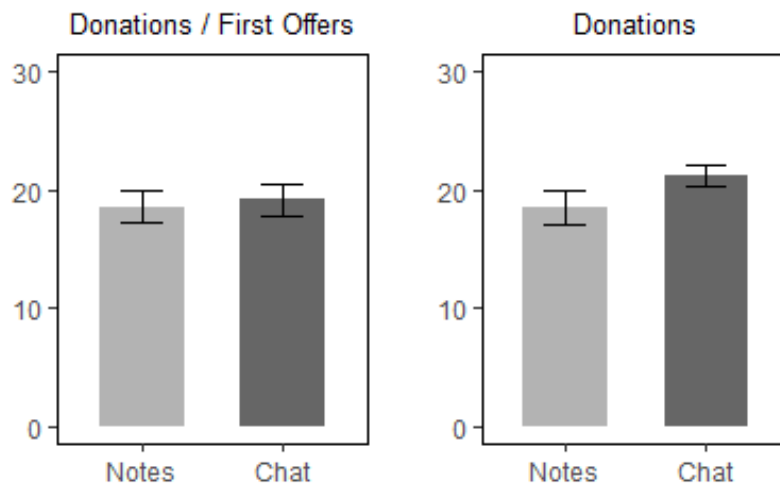


Figure 2.2 First donation suggestions and final donations

In the following, we investigate three potential mechanisms: first, an a priori social norm leading to higher willingness to donate even before the chat interaction started, second, an opinion-expression norm causing higher donations because of the exchanged messages, third, overly optimistic beliefs about partner’s refugee attitudes leading to higher donations.

2.4.2 A Priori Social Norm

We explore whether the mere fact of coming into contact with another person during the decision-making process influences the donation behavior. Engaging with a co-player, as opposed to deciding individually, might potentially enhance generosity, regardless of the specific content of the interaction. This would indicate a social norm to help the disadvantaged which unfolds its effect even before the interaction starts. In the left panel of Figure 2.2, we compare the very first offers of chat participants with the donation of *Notes* participants. To obtain the *first offer* data, we manually label all messages in the chat that contain a donation amount. Based on this, we label the first of such messages in every chat as *first offer*. In total, in 156 of 157 chat groups a first suggestion of an amount to donate was made. The average message number in which the first offer appears is 5.6, indicating that these offers were made at the very beginning of the conversations. We see that there is no difference between first offers under *Chat* and donations under *Notes* (MWU-tests, p-values: 0.505), indicating that it must be the chat interaction itself that convinced participants to donate more and not a social norm that was present already before the chat interaction started.

2.4.3 Opinion-Expression Norm

An opinion-expression norm during the chat interaction may explain our finding of higher donations to refugees in *Chat*. One potential norm is not to speak ill of a disadvantaged group, i.e., the refugee minors. Such a norm could have a profound effect on our main outcome variable, generosity towards refugees. While the channel in the former subsection refers to changes in generosity due to a potential social norm activated by being in contact with another subject, the channel in this subsection examines a social norm of speech, which subsequently influences behavior.

We test whether participants with negative attitudes toward refugees refrain more strongly from fully expressing these attitudes in *Chat*, compared to *Notes*. To this purpose, we

have coders classifying each message in *Chat* and *Notes* as expressing either a positive, negative, or neutral attitude toward refugees. Based on their coding, we construct two measures that characterize how positive, respectively negative, a given participant writes about refugees.

To be precise, two coders manually and independently labeled each text message from both *Notes* and *Chat* as *Pro* or as *Contra*. We define *Pro* (*Contra*) in a broad sense. Each message expressing a positive (negative) attitude, feeling, or opinion towards refugees is labeled as *Pro* (*Contra*). Messages that argue in favor of donating more (less) are also labeled as *Pro* (*Contra*). We argue that given the nature of our setting, the expressed attitudes towards refugees and the expressed willingness to donate are inseparable. We only used annotated messages where both coders agreed on the labeling and discarded the rest. Krippendorff's alpha for *Pro* and *Contra* are 0.8 and 0.69, indicating substantial agreement among coders. We find that overall, 610 messages in *Notes* contain expressed attitudes toward refugees; 419 (191) of those are positive (negative) attitudes. In *Chat*, 474 messages contain expressed attitudes, of which 373 (101) are positive (negative). Using these annotated text data, we construct the following variables:

$$Positivity_i = \begin{cases} \frac{Pro_i}{Pro_i + Contra_i} & \text{if } Pro_i + Contra_i > 0, \\ 0 & \text{if } Pro_i + Contra_i = 0, \end{cases} \quad (2.1)$$

$$Negativity_i = \begin{cases} \frac{Contra_i}{Pro_i + Contra_i} & \text{if } Pro_i + Contra_i > 0, \\ 0 & \text{if } Pro_i + Contra_i = 0, \end{cases} \quad (2.2)$$

where i is one chat interaction between two matched subjects or one individual note from subjects in *Notes*. These variables capture the positive and negative attitudes toward (donating to) refugee minors as expressed in *Notes* or in *Chat*. Spearman's correlation between $Positivity_i$ and $Negativity_i$ is -0.779, i.e., the two variables are not perfectly correlated.

In the left panel of Figure 2.3, we can observe that positive attitudes expressed in *Chat*

do not differ systematically from those expressed in *Notes* (MWU-test, p-value: 0.288). Considering negative attitudes, however, the right panel of Figure 2.3 illustrates that, on average, significantly fewer negative attitudes are expressed in *Chat* than in *Notes* (MWU-test, p-value: 0.000).

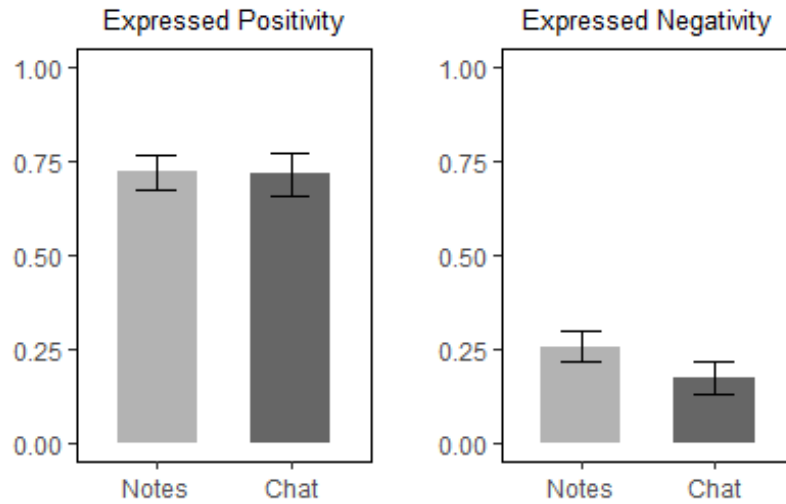


Figure 2.3 Expressed attitudes towards refugees

Moreover, within *Chat* we find that heterogeneity in the number of positive messages towards refugees is higher than those for negative messages. Figure 2.4 illustrates the number of positive (negative) messages per participant. Almost 80% of participants did not express negative attitudes towards refugees in the chat.

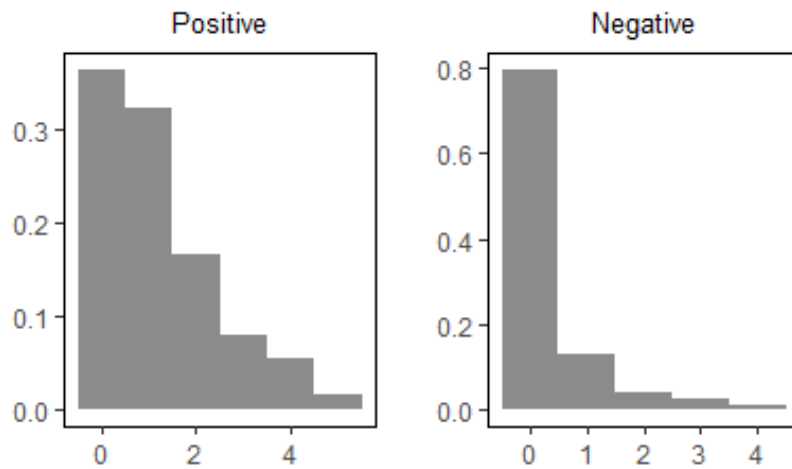


Figure 2.4 Number of positive/negative messages in the chat (x-axis) and share of subjects expressing them (y-axis)

Our finding that negative attitudes are expressed less often during the chat discussion is corroborated by a simple sentiment analysis that measures the general tone of the chat interaction or personal notes. We use the predefined dictionary for the German language developed by Rauh (2018), that classifies words as positive, negative, or neutral. While Remus, Quasthoff, and Heyer (2010) developed a dictionary that set the standard for sentiment analysis for the German language, Rauh (2018) refines this measure by taking into account the negation of words, that reverses the sentiment of an expression. Our results, however, do not depend on the choice of the dictionary.² We then calculate positive (negative) sentiment as the number of positive (negative) words divided by all words in a given chat group or note. Theoretically, these measures can range from -1 to +1. However, naturally occurring language often produces values close to 0. Rauh (2018), e.g., illustrates this by analyzing a sample of 1500 sentences from the German Parliament and showing that 31.6% are classified as neutral. A sentiment of -0.1 can be interpreted as a 10% overweight of negatively connoted language, suggesting a negative sentiment prevailing in the corresponding chat group or note. Results, depicted in Figure 2.5, show that *Notes* contains significantly more negative words than *Chat* and the latter contains significantly more positive words (MWU-test, p-values: 0.000 and 0.000, respectively).

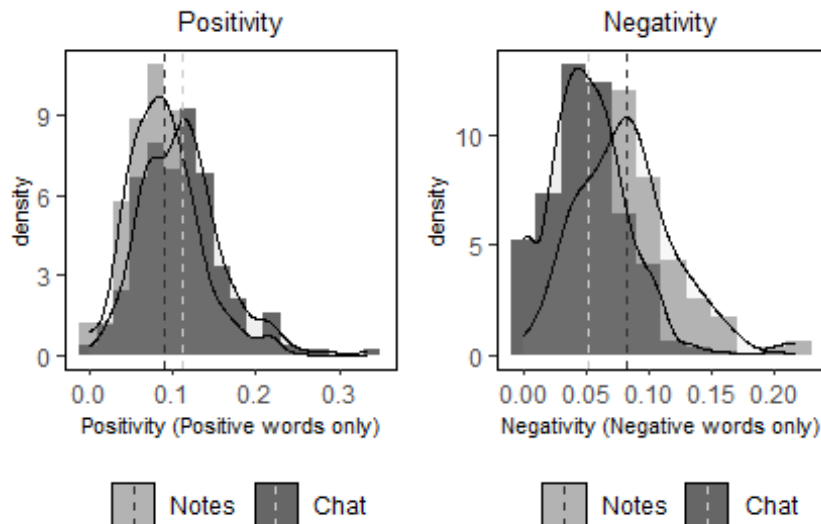


Figure 2.5 Sentiment - positivity and negativity by treatment

We interpret these findings as follows. Since ex-ante attitudes towards refugees do not

²Results using the Remus, Quasthoff, and Heyer (2010) measure are available upon request.

differ across treatments, our text-data analysis suggests self-censorship in the chat treatment: Subjects hold back some negative attitudes when communicating with another subject. As a result, subjects are exposed to fewer anti-donation arguments and to an overall donation-friendly communication. We argue that this is a potential channel ultimately impacting behavior.

2.4.4 Overly Optimistic Beliefs About Partner's Refugee Attitudes

Communication and the opinion-expression norm may not only directly cause higher donations, but also influence the beliefs about the partner's attitudes towards refugees. Overly optimistic beliefs could in turn increase donations, to match what the subject believes to be the peer's generosity. Figure 2.6 illustrates the (accuracy) of beliefs about the partner's attitudes towards refugees. It depicts the difference between a subject's incentivized belief about the partner's attitudes that we elicited both before and after treatment and the partner's actual attitudes (*Perceived – Actual*). We see that belief errors in all four cases are positive indicating that on average subjects believe their partners to have more positive attitudes toward refugees than they actually have. Belief errors are not significantly different across treatments, neither before nor after treatment (MWU-test, p-values: 0.339 and 0.451, respectively). Hence, subjects in *Chat* do not seem to learn significantly about the actual attitudes of their partners. In particular, they do not tune down their excess optimism about their partners' refugee-friendliness. This suggests that the opinion-expression norm according to which subjects hold back some negative attitudes toward refugees in the chat is successful in preserving this optimism.

The opinion-expression norm preserves (overly) optimistic beliefs about the chat partner's attitudes toward refugees. Since these attitudes and the willingness to donate are correlated (Spearman's correlation coefficient: 0.31), it is plausible to assume that the chat also creates more optimistic beliefs about the partner's willingness to donate. As the existing literature demonstrates, such beliefs may directly stimulate generosity. For

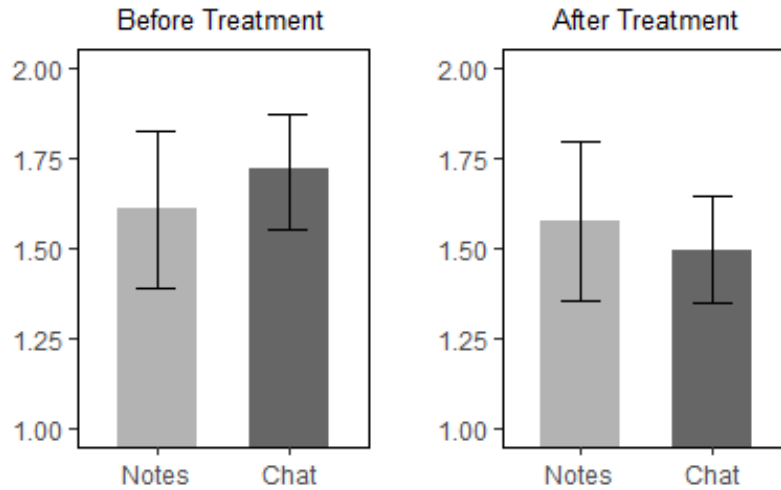


Figure 2.6 Belief error regarding chat-partner's attitudes towards refugees

instance, Shang and Croson (2009) provide evidence that when individuals receive information about others donating large amounts, they tend to increase their own donations.

We have discussed three channels through which communication might affect donations. We do not find any evidence for a priori social image concerns. In other words, it is the content of the interaction rather than the mere fact of being paired with a stranger that drives donations. We have provided some evidence for how the chat is characterized by an opinion-expression norm meaning that subjects hold back some negative arguments and attitudes. This shields negative attitudes toward refugees from perception, thereby creating a donation-friendly communication. Further, this norm is associated with overly optimistic beliefs about the partners' willingness to donate, which promotes donations. Our work indicates how a social norm of speech (i.e., not to speak ill of refugees) can translate into real behavior rather than being mere cheap talk.

Moreover, we find no effect of any of the video primes on donations to refugees. Average donations for those that saw a positive, negative, and balanced video are 20.65, 20.32, and 19.86 Taler, respectively (MWU-tests, p-values: 0.99, 0.99, and 0.99). Thus, the videos did not influence donation behavior.

2.4.5 Indirect Effect of *Chat* on Trust

Besides our main finding that the chat interaction stimulates higher donations, we also investigate the effect of the chat interaction on trust between the two chat participants (i.e., trust towards the in-group). We define the change in trust as the difference in the amount sent during the second trust game minus the amount sent in the first trust game (denoted as $\Delta Trust$). The distribution of $\Delta Trust$, depicted in Figure 2.7, illustrates that half of all subjects (244 subjects) change their level of trust towards the co-player in the second trust game.

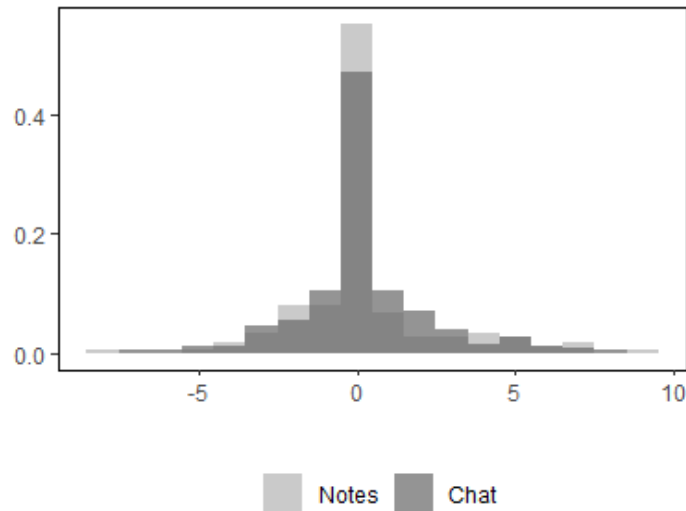


Figure 2.7 Distribution of $\Delta Trust$

In *Notes*, such changes cannot be rationalized based on new information about the matched partner. In *Chat*, by contrast, subjects have the opportunity to learn new information from their partner. In particular, they can learn about their partner's attitude toward refugees - if partners are at least partially open about their attitudes. Pre-vote deliberation may impact trust, as subjects could infer their partner's trustworthiness based on the arguments, opinions, and preferences presented. We, therefore, investigate if *Chat* has a direct effect on trust and if there is an indirect effect via changes in beliefs about the co-player's refugee attitudes. Results of a parametric mediation analysis are summarized in Figure 2.8.

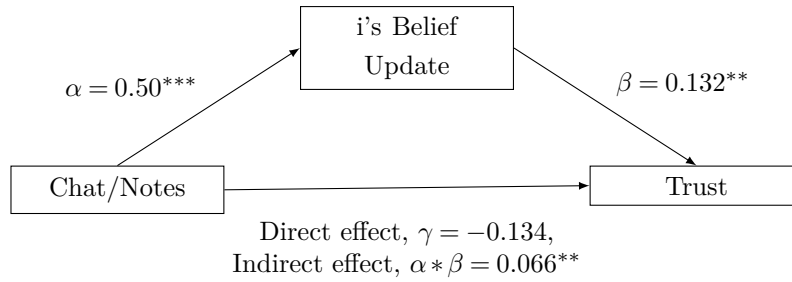


Figure 2.8 Chat affects trust indirectly through belief update

As the estimated coefficient γ shows, the chat interaction *per se* does not significantly change a participant's trust in their chat partners ($\gamma=-0.134$). However, the stronger participants (positively) update their beliefs about the partners' attitudes towards refugees, the more they trust them.³ Being assigned to the *Chat* rather than to the *Notes* treatment leads to a significant positive belief update by about half a Likert-scale point ($\alpha=0.50$). That belief update in turn leads to an increase in trust by $\beta=0.132$ ECU. While this effect is rather small in size, it is a statistically significant indirect effect. The appendix presents details about the methodology we used to identify the mediation effect and a sensitivity analysis.

With regard to mediation analysis, Zhao, Lynch, and Chen (2010) clarify a common misunderstanding: There is no need for a significant direct effect for an indirect effect to be valid (cf. p. 199). The authors further classify different mediation types according to the significance levels and signs of the coefficients α , β , and γ . Following their classification, our results illustrated in Figure 2.8 are termed *indirect-only mediation*. As Zhao, Lynch, and Chen (2010) write, this mediation type indicates evidence for the hypothesized mediator, i.e., chat communication affects trust through changes in beliefs about the co-players' refugee attitudes, and other mediation effects are unlikely.

In sum, interacting in the chat, on average, causes a (positive) update in beliefs regarding the co-player's attitudes towards refugees. This belief update in turn makes participants trust their co-player more.⁴

³Participant *i*'s belief update is defined as the difference between their belief (regarding the co-player's refugee attitudes) before and after the chat interaction (see appendix for details).

⁴This behavior is only rational if an individual's refugee-friendliness functions as a credible signal of trustworthiness towards the co-player. We do indeed find some evidence that this is the case. In our sample, subjects expressing more negative attitudes towards refugees send back smaller amounts in

2.5 Discussion

The paper’s finding that communication among peers increases donations relates to the literature on how social information affects donation behavior and how this can be used by fundraising campaigns to nudge people to contribute more (e.g., Frey and Meier 2004, Martin and Randal 2008, Croson and Shang 2008, Shang and Croson 2009). The size of our documented effect (about 14%) is quite similar to Shang and Croson (2009), who find an effect of about 12% in their most effective condition. While most of this literature exogenously provides information on past behavior of other donors, in our setting, however, donors communicate their donation intent directly via the chat interaction. With this direct communication channel, they can also provide reasons for their donation decisions. This allows us to study a richer setting in which social information affects donation behavior. Closest to our experimental setup is the work by Reyniers and Bhalla (2013) who compare individual versus collective donation decisions in an offline survey experiment and find that donations are higher in the collective treatment. In Reyniers and Bhalla (2013) participants have the choice between five different charities, while our work does not offer such a choice. Therefore, the discussions in our experiment focus on the pros and cons of giving to that particular charity, not on preferences between charities. Moreover, we use text-analysis methods to identify potential mechanisms that mediate the effect that the discussion has on generosity.

Our paper also contributes to the discussion on the importance of trust in economic decision-making and as a determinant of a country’s economic success.⁵ More specifically, we contribute to the literature investigating generosity and charitable giving as a signal of trustworthiness (Gambetta and Przepiorka 2014 and Fehrler and Przepiorka 2016). Gambetta and Przepiorka (2014) find that individuals who are more generous to-

the trust game. The correlation between an individual’s positive (negative) attitudes towards refugees and their trustworthiness in the first trust game is 0.083 (-0.150). We investigate this more deeply by regressing $\Delta Trustworthiness$ on an individual’s attitudes towards refugees (see appendix for details).

⁵This well-established literature includes: Fukuyama (1995); La Porta et al. (1997); Knack and Keefer (1997); Leonardi, Nanetti, and Putnam (2001); Zak and Knack (2001); Guiso, Sapienza, and Zingales (2004, 2008, 2009); Bloom, Sadun, and Van Reenen (2012); Algan and Cahuc (2014); Butler, Giuliano, and Guiso (2016); Bolton and Chen (2018).

wards the co-player in a dictator game are perceived as more trustworthy in a subsequent trust game. In comparison, our work examines a setting in which generosity towards an out-group (rather than towards the co-player) can potentially generate higher trust towards the co-player. Fehrler and Przepiorka (2016) find that individuals who decide to give to a charity are perceived as more trustworthy and are more often selected as interaction partners. Thus, charitable giving in this context is indeed perceived as a valuable signal of trustworthiness. Complementing these studies, our paper highlights a specific channel through which generosity affects being perceived as trustworthiness: A person communicates their willingness to contribute to charity and their reasons for doing so. This affects others' beliefs about their attitudes toward a disadvantaged out-group. This, in turn, leads to an increase in trust in that person.

The study's limitations can be summarized as follows. The generalizability of the results is restricted. Our sample is self-selected and relatively homogeneous, comprising school minors from the two largest cities in Germany, mostly born in Germany, and from schools that offer "Abitur" (the German education qualification that qualifies students for university admission). In a more heterogeneous sample comprising a larger variety of school types, the effects of communication might have been different. However, given our homogeneous social group, we find evidence that this group coordinates on a norm of opinion expression that leads to higher donations.

2.6 Conclusion

Our study investigates the effects of in-group communication on generosity towards an out-group. In our lab-in-the-field experiment involving school minors, communication with a partner results in increased transfers to refugee minors. Further, we observe an opinion-expression norm in the chat treatment, where participants withhold negative attitudes towards refugees. Subsequently, communication in our experiment creates overly optimistic beliefs about the partner's attitude toward refugees. We also document an indirect effect of the chat treatment on in-group trust. Our work shows how communication

can influence real behavior rather than being mere cheap talk. It is crucial to note that the emergence of communication dynamics, such as the opinion-expression norm in our study, is highly context and topic-specific. Ultimately, our experimental insights contribute to the existing literature on communication, generosity, and trust.

Acknowledgments

This paper benefited from discussions with Claudia Schwirplies. We are grateful to Sophia Schulze-Schleithoff and Jan-Patrick Mayer for excellent research assistance and to the University of Hamburg WISO lab (particularly Ziad Zorgati, David Lucius, Katharina Groß, Jana Lücke, and Barabros Saritos) for outstanding technical assistance. This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 822590. Any dissemination of results here presented reflects only the authors' view. The Agency is not responsible for any use that may be made of the information it contains. IRB approval for this research was given at the University of Hamburg on the 19th of February 2019.

Bibliography

- Algan, Yann and Pierre Cahuc. 2014. “Trust, growth, and well-being: New evidence and policy implications.” In *Handbook of Economic Growth*. 49–120.
- Bereczkei, Tamas, Bela Birkas, and Zsuzsanna Kerekes. 2010. “Altruism towards strangers in need: Costly signaling in an industrial society.” *Evolution and Human Behavior* 31:95–103.
- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. “Trust, reciprocity, and social history.” *Games and Economic Behavior* 10 (1):122–142.
- Bhagal, Manpal S. 2021. “Altruism advertises cooperativeness.” In *Encyclopedia of Evolutionary Psychological Science*. 240–242.
- Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2012. “The organization of firms across countries.” *The Quarterly Journal of Economics* 127 (4):1663–1705.
- Bolton, Gary E. and Yefen Chen. 2018. “Other-regarding behavior: Fairness, reciprocity, and trust.” In *The Handbook of Behavioral Operations*. 199–235.
- Butler, Jeffrey V., Paola Giuliano, and Luigi Guiso. 2016. “The right amount of trust.” *Journal of the European Economic Association* 14 (5):1155–1180.
- 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.
- Croson, Rachel and Jen Shang. 2008. “The impact of downward social information on contribution decisions.” *Experimental Economics* 11 (3):221–233.
- Fehrler, Sebastian and Wojtek Przepiorka. 2016. “Choosing a partner for social exchange: Charitable giving as a signal of trustworthiness.” *Journal of Economic Behavior and Organization* 129:157–171.

- Frey, Bruno S. and Stephan Meier. 2004. "Social comparisons and pro-social behavior: Testing 'conditional cooperation' in a field experiment." *American Economic Review* 94 (5):1717–1722.
- Fukuyama, Francis. 1995. *Trust: The social virtues and the creation of prosperity*. Free Press.
- Gambetta, Diego and Wojtek Przepiorka. 2014. "Natural and strategic generosity as signals of trustworthiness." *PloS one* 9 (5):e97533.
- Gibbard, Allan. 1977. "Manipulation of schemes that mix voting with chance." *Econometrica* 45 (3):665–681.
- Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2004. "The role of social capital in financial development." *American Economic Review* 94 (3):526–556.
- . 2008. "Trusting the stock market." *The Journal of Finance* 63 (6):2557–2600.
- . 2009. "Cultural biases in economic exchange?" *The Quarterly Journal of Economics* 124 (3):1095–1131.
- Holm, Sture. 1979. "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* 6 (2):65–70.
- Huber, Martin. 2020. "Mediation analysis." In *Handbook of Labor, Human Resources and Population Economics*. 1–38.
- Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010a. "A general approach to causal mediation analysis." *Psychological Methods* 15 (4):309–334.
- . 2010b. "Identification, inference, and sensitivity analysis for causal mediation effects." *Statistical Science* 25 (1):51–71.
- Knack, Stephen and Philip Keefer. 1997. "Does social capital have an economic payoff? A cross-country investigation." *The Quarterly Journal of Economics* 112 (4):1251–1288.

- Kranton, Rachel, Matthew Pease, Seth Sanders, and Scott Huettel. 2020. “Deconstructing bias in social preferences reveals groupy and not-groupy behavior.” *Proceedings of the National Academy of Sciences* 117 (35):21185–21193.
- La Porta, Rafael, Florencio Lopez-de Silanes, Andrei Shleifer, and Robert W. Vishny. 1997. “Trust in large organizations.” *American Economic Review* 87 (2):333–338.
- Leonardi, Robert, Raffaella Y. Nanetti, and Robert D. Putnam. 2001. *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.
- Martin, Richard and John Randal. 2008. “How is donation behaviour affected by the donations of others?” *Journal of Economic Behavior & Organization* 67 (1):228–238.
- Milinski, Manfred, Dirk Semmann, and Hans-Jürgen Krambeck. 2002. “Donors to charity gain in both indirect reciprocity and political reputation.” *Proceedings: Biological Sciences* 269:881–883.
- Mokos, Judit and István Scheuring. 2019. “Altruism, costly signaling, and withholding information in a sport charity campaign.” *Evolution, Mind and Behaviour* 17 (1):10–18.
- Nelson, Jacob L., Dan A. Lewis, and Ryan Lei. 2017. “Digital democracy in America: A look at civic engagement in an internet age.” *Journalism & Mass Communication Quarterly* 94 (1):318–334.
- Rauh, Christian. 2018. “Validating a sentiment dictionary for German political language – a workbench note.” *Journal of Information Technology & Politics* 15 (4):319–343.
- Remus, Robert, Uwe Quasthoff, and Gerhard Heyer. 2010. “SentiWS – A publicly available German-language resource for sentiment analysis.” *Proceedings of the Seventh International Conference on Language Resources and Evaluation* :1168–1171.
- Reyniers, Diane and Richa Bhalla. 2013. “Reluctant altruism and peer pressure in charitable giving.” *Judgment and Decision Making* 8 (1):7–15.
- Shang, Jen and Rachel Croson. 2009. “A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods.” *The Economic Journal* 119 (540):1422–1439.

- Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Kosuke Imai, and Luke Keele. 2014. “Mediation: R package for causal mediation analysis.” *Journal of Statistical Software* 59 (5):1–38.
- VanderWeele, Tyler. 2016. “Mediation analysis: A practitioner’s guide.” *Annual Review of Public Health* 37:17–32.
- Wedekind, Claus and Victoria A. Braithwaite. 2002. “The long-term benefits of human generosity in indirect reciprocity.” *Current Biology* 12:1012–1015.
- Zak, Paul J. and Stephen Knack. 2001. “Trust and growth.” *The Economic Journal* 111 (470):295–321.
- Zhao, Xinshu, John Lynch, and Qimei Chen. 2010. “Reconsidering Baron and Kenny: Myths and truths about mediation analysis.” *The Journal of Consumer Research* 37 (2):197–206.

Appendix A

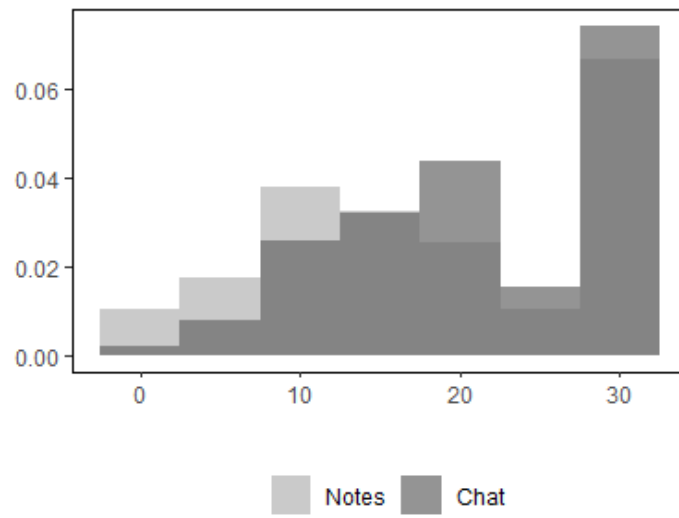


Figure 2.9 Distribution of *Donation to Refugees*

Appendix B

Table 2.2 Tobit regressions: donations

	(1)	(2)	(3)	(4)
Chat	3.232** (1.446)	3.038** (1.345)	3.814*** (1.259)	3.565*** (1.189)
Attitudes Refugees			1.877*** (0.305)	1.539*** (0.289)
Negative Video			-0.105 (1.312)	-0.231 (1.259)
Balanced Video			-0.231 (1.341)	-0.137 (1.233)
Female			3.126*** (1.175)	3.480*** (1.091)
Age			1.090* (0.651)	0.962 (0.731)
Born in Germany			3.843 (2.856)	2.926 (2.492)
Pocket money			-0.003 (0.013)	0.001 (0.012)
Household members			-0.286 (0.448)	-0.251 (0.441)
Donation to MSF			0.546*** (0.085)	0.564*** (0.080)
December			-0.325 (1.691)	-3.256 (3.726)
Constant	21.030*** (1.110)	25.632*** (2.703)	-21.923* (11.909)	-12.615 (14.757)
School FE	No	Yes	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
Obs.	487	487	482	482
Wald Test	6.645***	86.780***	129.196***	202.513***

Notes: The table reports results of Tobit regressions with *Donation to Refugees* as the dependent variable. Since many subjects donate the maximum amount of 30 (see Figure 2.9 in the appendix), we use Tobit regressions to estimate the effect of Chat on the latent unrestricted donation. The variable *Chat* is a dummy equal to one for subjects that chatted and zero otherwise. The variable *Attitudes Refugees* is a categorical variable indicating positive/negative attitudes towards refugees. The variable *Negative Video* (*Balanced Video*) are dummies that are equal to one for subjects that saw a negative (balanced) video about refugee minors and zero otherwise. The variable *Female* is a dummy equal to one for female subjects and zero otherwise. The variable *Age* is numeric ranging from 15 to 21. *Born in Germany* is a dummy that is equal to one for subjects born in Germany and zero otherwise. *Pocket money* is numeric stating the money subjects receive from their parents. The variable *Household members* indicates a subject's number of household members. The variable *Donation to MSF* indicates a subject's donation amount during the survey's lottery for donating to MSF. Finally, *December* is a dummy equal to one for sessions conducted in December and zero otherwise. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Appendix C

Instructions Chat and Notes We advise the subjects to start the discussion by writing a few words about themselves (without revealing their identity). We expect this feature to reduce social distance and to make constructive deliberation more likely. However, this personal introduction can happen in any form and is not enforced by the experimenter in any way. After this short introduction, we advise the subjects to concentrate the discussion on the donation matter and voting decision. Furthermore, we ask the subjects to provide arguments for or against donating to refugees. This allows us to analyze if expressing opinions and arguments influences the deliberation and subsequently the vote on transfers.

Constructing Δ Opinion co-player In the questionnaire, we ask participants to rate the extent to which they think their co-players have positive (negative) attitudes toward refugees. The answer options to this question are *1 = very strongly, rather strongly, neither strongly nor weakly, rather weakly, 5 = weakly or not at all pronounced*.

For our analysis, we combine the data from the two questions by first reversing the answers for the question regarding positive attitudes (“1” is recoded as “5” etc.). We then add up the answers regarding the positive attitudes (now reversed) and the answers regarding the negative attitudes. We implement this for the answers before the chat and the answers after the chat. Finally, we subtract the resulting value before the chat from the resulting value after the chat to obtain Δ Opinion co-player. Figure 2.10 depicts its distribution per treatment. Δ Opinion co-player ranges from -5 to 6.

Robustness analysis Δ Opinion co-player Our results remain unchanged if we consider the positive attitudes instead of the combined measure: The average change is 0.058 in *Notes* and 0.323 in *Chat* (MWU-test: p-value=0.000). The same story holds for negative attitudes instead of the combined measure: The average change is -0.092 in *Notes* and 0.185 in *Chat* (MWU-test, p-value: 0.000). This impression is also confirmed

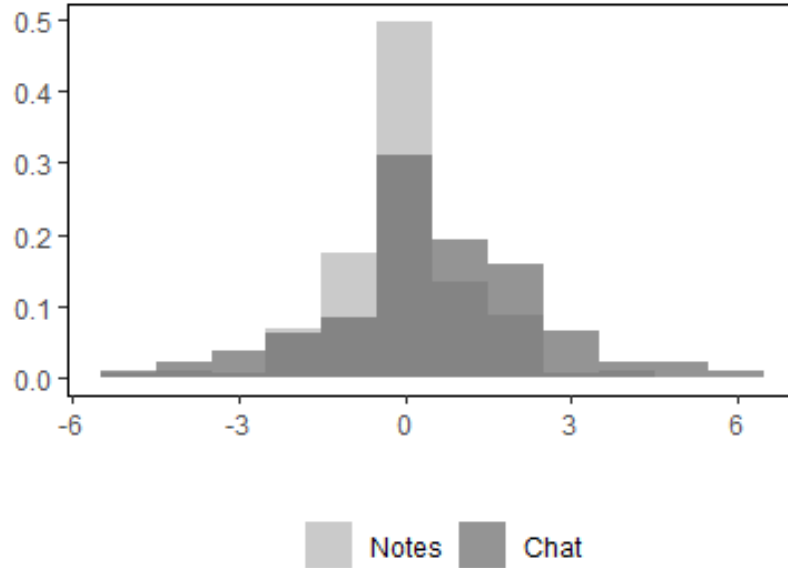


Figure 2.10 Change in beliefs about co-player - distribution

when examining a Wilcoxon-signed-rank test for those who chatted in order to compare the beliefs before and after the chat. The p-value is 0.000 for the positive as well as the negative attitudes.

Appendix D

Ex-ante power calculation Before conducting our lab-in-the-field experiment in schools, we perform the following power calculation in order to get an estimate for the efficient sample size. We make the following assumptions about the minimum relevant distance (MRD) between treatments and the common standard deviation (SD) of our outcome variables across treatments.

First, we define the MRD between our treatments *Chat* and *Notes* that is economically relevant. Our two outcome variables are (a) *Donation to Refugees*, where subjects can donate between 0 and 30 ECU, and (b) $\Delta trust$. The latter is defined as the difference in ECU sent between the two trust games that are played. In each trust game, subjects can send between 0 and 10 ECU to their co-player. In both cases, we want to detect an MRD of at least 10% of the initial endowment between *Chat* and *Notes*. For *Donation to Refugees*,

this means we want to detect at least a mean difference of 3 ECU between treatments. For $\Delta trust$, the MRD is 1 ECU. This corresponds to a real monetary difference of 1.80 and 0.60 Euro, respectively.

Second, we need an estimate for the common standard deviation (SD) of these outcome variables across treatments. In both cases, we rely on the standard deviation of these variables from our pilot study that we conducted with seventeen participants in the WISO lab at the University of Hamburg. This results in estimates for the standard deviation of 8.82 for *Donation to Refugees* and 2.67 for $\Delta trust$. For comparison, the standard deviations of *Donation to Refugees* and $\Delta trust$ in our final sample are 8.86 and 2.2, indicating that our estimates from the pilot study were quite accurate. The effect size (ES) is calculated as MRD/SD . Results are summarized in Table 2.3.

Table 2.3 Efficient sample size calculation

Outcome variable	MRD	Exp. SD	ES	Efficient sample size
<i>Donation to Refugees</i>	3 ECU	8.82	0.34	166
$\Delta trust$	1 ECU	2.67	0.37	141

Notes: The table displays the power calculation for our two outcome variables *Donation to Refugees* and $\Delta trust$. We assume a minimum relevant distance (MRD) of 10% (of the initial endowment) between treatments, i.e., 3 ECU for *Donation to Refugees* and 1 ECU for $\Delta trust$, respectively. The expected standard deviations (SD) of the outcome variables are taken from the pilot study. The effect size (ES) is MRD/SD . For the efficient sample size (per treatment), we correct for two hypotheses being tested, i.e., the effect of *Chat* on *Donation to Refugees* and $\Delta trust$.

The efficient sample size per treatment, i.e., for *Chat* and *Notes*, is 166 observations in the case of *Donation to Refugees* and 141 observations in the case of $\Delta trust$. We assume a power of 0.8 (commonly used for field experiments) and an alpha of 0.05. We correct alpha for two hypotheses we want to test, i.e., the effect of *Chat* on trust and donation behavior. That means we divide alpha by two.

Overall, we conclude that for the MRD we want to detect, i.e., a 10% difference between treatments, we need at least 166 observations per treatment, i.e., *Chat* and *Notes*. Since observations in *Chat* are not independent due to interactions among two subjects, we plan twice the minimal sample size of 166 observations for *Chat*. Overall, we plan a sample size of $3 * 166 = 498$, i.e., approximately 500 observations.

Appendix E

Causal mediation analysis In this section, we provide details on the causal mediation analysis that tests for the potential causal effect of the chat treatment on trust through changes in belief about the co-player’s refugee attitudes (for an illustration see Figure 2.11). The corresponding variables of interest are: *Chat* and *Notes* as the treatment variable, $\Delta Trust$ as the outcome variable of interest, and $\Delta Opinion\ co-player$ as the mediator indicating how much an individual has changed their belief about the co-player’s attitudes towards refugees.

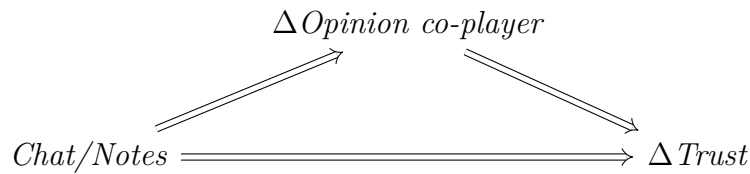


Figure 2.11 Potential mediation - trust

We identify a causal mediation effect based on the following four assumptions: There must not be confounders between treatment and outcome relationship (assumption 1), between mediator and outcome relationship (assumption 2), or between treatment and mediator relationship (assumption 3). In addition, there must not be confounders affected by the treatment between mediator and outcome relationship (assumption 4). These assumptions are also known as sequential ignorability or sequential independence assumptions (Huber 2020).

Assumptions 1 and 3 are met because our treatment is randomized. For Assumption 2 we have to consider all post-treatment potential confounders that may affect the path from the mediator to the outcome. As individuals have the chance to update their belief about their co-player’s attitudes towards refugees directly after the chat and this is instantaneously followed by the second trust game, there is little reason to believe that a post-treatment confounder would dilute the path from the treatment through the mediator on changes in trust. Similarly, assumption 4 is more plausible, the less time that

has elapsed between the treatment and the mediator (VanderWeele 2016). In our case, the question about the co-player’s attitudes towards refugees occurs directly after the treatment, i.e., *Chat/Notes*. Under these sequential ignorability assumptions, a causal mediation from the treatment to the outcome variable can be established.

In the following, we empirically investigate if such a mediation effect exists, using the methods proposed in Imai, Keele, and Tingley (2010a) and Imai, Keele, and Tingley (2010b). We use the *mediation* package in R (Tingley et al. 2014) to implement this analysis. As a first step, we formulate the outcome and mediator model as

$$\Delta Trust_i = \zeta + \gamma Chat_i + \beta \Delta Opinion\ co-player_i + \theta X_i + \epsilon_i \quad (2.3)$$

$$\Delta Opinion\ co-player_i = \lambda + \alpha Chat_i + \phi X_i + \eta_i \quad (2.4)$$

The variable $Chat_i$ is equal to one if a subject was assigned to the chat treatment and zero otherwise. The matrix X_i contains the following control variables: *Female* is a dummy equal to one for female subjects and zero otherwise. *Age* is numeric ranging from 15 to 21. *Household members* indicates a subject’s number of household members. *Risk aversion* is a 5-point Likert scale reflecting subjects perceived risk attitudes. *Lottery bet* reflects a subject’s risk attitudes by measuring their willingness to pay for a lottery ticket that has a 50% chance of winning 300 Euro. The variable *Sentiment co-player* is the number of positive words minus negative words normalized by the total amount of words written by the co-player. Finally, *School ID* are dummy variables that are used as school fixed-effects.

Results of this mediation analysis are presented in Table 2.4. The ACME (Average causal mediation effect) is significant. This means that the chat interaction exhibits a significant indirect effect on $\Delta Trust$ via the mediator $\Delta Opinion\ co-player$. The ADE (Average direct effect), however, is not significant, i.e., there is no direct effect of the treatment on $\Delta Trust$. Thus, the chat interaction does not per se affect changes in trust among subjects but only via the belief update about the co-player’s attitudes towards refugees.

Sensitivity analysis We perform a sensitivity analysis, which allows us to assess how

Table 2.4 Causal mediation analysis

Effect	Estimate	CI lower	CI upper	p-value
ACME ($\alpha*\beta$)	0.07	0.01	0.15	0.018**
ADE (γ)	-0.13	-0.55	0.27	0.578
Total Effect	-0.07	-0.47	0.35	0.812
Prop. Mediated	-0.97	-4.21	3.69	0.814

Notes: While ACME abbreviates Average Causal Mediation Effect, ADE abbreviates Average Direct Effect. Confidence intervals are obtained with a nonparametric bootstrap using the percentile method. Sample size: 481 (7 subjects are removed due to missing values in the control variables). Simulations: 1000. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

robust our indirect effect estimates are to violations in the sequential ignorability assumptions and how substantial a violation in the assumptions would have to be in order to considerably alter our inferences about the indirect effect (ACME). The basic idea of the sensitivity analysis is to study the correlation ρ of the errors of both models (ϵ and η). Under sequential ignorability, ρ is equal to zero and thus the magnitude of this correlation coefficient can represent the departure from the ignorability assumption. The correlation ρ for our outcome- and mediator model is close to zero (1.5975×10^{-16}). Assuming that the specifications of both models are correct, it seems there is no evidence for a violation of the assumptions. Results for potential departures of ρ and therefore violations of sequential ignorability are summarized in Figure 2.12.

The figure displays the Average Causal Mediation Effect (ACME) as a function of ρ that simulates potential violations of sequential ignorability. Only for $\rho > 0.1$, the ACME would be zero or change its sign, i.e., our conclusions would not be valid. In other words, the correlation of error terms between the outcome and mediator model would need to be quite large in order to draw different conclusions about the mediation effect.

Appendix F

Attitudes towards refugees and Δ trustworthiness Columns 1, 3, and 5 in Table 2.5 show the effect of individual attitudes towards refugees on Δ trustworthiness without

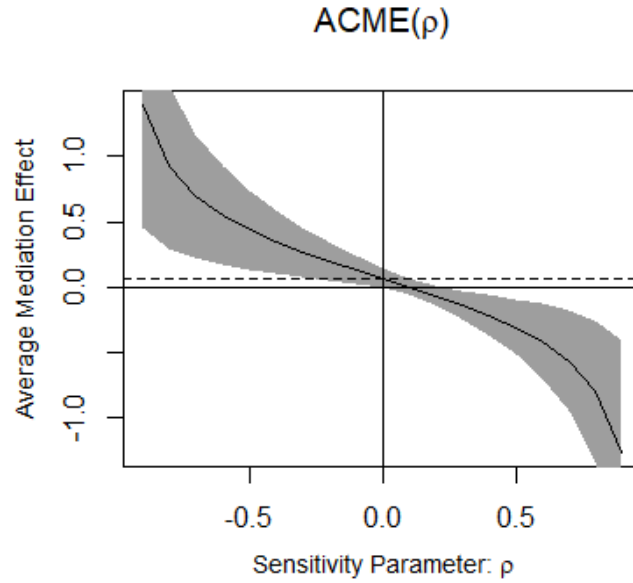


Figure 2.12 Sensitivity analysis

Table 2.5 Attitudes towards refugees and Δ trustworthiness

	(1)	(2)	(3)	(4)	(5)	(6)
Pos. attitude	0.046** (0.023)	0.054** (0.026)				
Neg. attitude			-0.061*** (0.022)	-0.066*** (0.024)		
Comb. attitudes					0.032*** (0.012)	0.036*** (0.014)
Constant	1.093*** (0.086)	1.153** (0.489)	1.390*** (0.049)	1.420*** (0.460)	1.019*** (0.094)	1.034** (0.492)
Obs.	488	485	488	485	488	485
Control Variables	No	Yes	No	Yes	No	Yes
School FE	No	Yes	No	Yes	No	Yes
R ²	0.010	0.041	0.019	0.047	0.017	0.047
F Statistic	5.134**	0.994	9.580***	1.145	8.612***	1.150

Note: The table reports OLS regressions. Δ trustworthiness is the dependent variable and measures the difference between *amount returned trust game 2* and *amount returned trust game 1*. *Pos. attitude* (*Neg. attitude*) is an individual's positive (negative) attitudes towards refugees stated at the beginning of the survey. The variable *Comb. attitude* combines positive and negative attitudes to one measure. Control variables are the same as described in Table 2.2. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

control variables while columns 2, 4, and 6 add controls. The findings highlight that Δ trustworthiness increases (decreases) significantly with positive (negative) attitudes towards refugees. With this analysis, we do not intend to suggest that there is a causal relationship between refugee-friendliness and Δ trustworthiness. Nor do we want to generalize this relationship beyond our sample. Importantly, generosity towards refugees comes at the expense of the common class fund in our setting. However, these correlations illustrate that trusting a partner who has signaled to be refugee-friendly can be rationalized in our experimental setting.

Appendix G

Experimental interfaces are available upon request.

Chapter 3

Truth-telling and Wishful Thinking

Author: Jan Biermann

Abstract

This study examines how people strategically manipulate their own beliefs about other people's behavior. It investigates whether breaking an honesty norm negatively affects what people think others would do in the same situation. I use a large-scale incentivized online experiment with a representative sample of the UK's general population. Following a novel experimental design the study exposes subjects to one of two conditions. While lying is possible in both cases, it is more tempting in one situation than in the other. The study finds that if lying is more tempting, the proportion of people who choose to lie increases. Importantly, subjects in this condition also hold more pessimistic beliefs about what others would do in the same situation. I incorporate an additional treatment dimension, in which subjects are not actively involved in the situation and, hence, don't have the opportunity to lie. Including the beliefs of these passive subjects in the analysis enables me to argue that the difference in beliefs is not due to a rational anticipation of the different incentive structures.

Keywords: Norm-following, lying, belief formation, motivated beliefs

JEL: C91, D91, D83, D84

3.1 Introduction

Subjective beliefs play a key role in economic decision-making. Various trends in the economics literature emphasize different facets of beliefs. Standard game theory frequently assumes agents to be perfectly rational. Behavioral economics highlights the role of heuristics and biases. The more recent literature on motivated beliefs (cf. Bénabou and Tirole, 2016) stresses that beliefs fulfill psychological functions such as reducing anxiety or maintaining self-esteem.

Prior research has demonstrated that individuals engage in self-serving belief manipulation in a variety of domains, which include beliefs about the quality of goods and services (Gneezy et al., 2020), an individual's own ability (Zimmermann, 2020), the definition of fairness (Konow, 2000), or whether a person's behavior harms others (Grossman and Van der Weele, 2017). An important, yet underexplored, context for self-serving belief manipulation concerns the assessment of the norm compliance of one's peers. In such situations, people form beliefs about whether others would break a social norm if they were to face the same temptation as those assessing.

One relevant strand of literature concerning the emergence of motivated beliefs in moral settings is the literature on “motivated Bayesians.” Gino, Norton, and Weber (2016) coined this term to describe how people process information in making moral decisions when they employ predictable cognitive patterns aimed at maintaining their sense of morality while simultaneously serving their self-interest. Bénabou and Tirole (2011) laid the theoretical groundwork for this body of literature. In their model, behavior fulfills a function of self-signaling according to which people draw inferences about their own moral motivation based on actions they observe themselves taking.¹ However, such self-signaling poses a dilemma since individuals desire to take actions they prefer while maintaining a positive self-image. One possibility of escaping this dilemma is to behave as a motivated Bayesian.

¹For related theoretical contributions see Bénabou and Tirole (2006) and Grossman and Van der Weele (2017).

A well-documented manifestation of biased information processing in this literature is that of avoiding some information altogether.² Numerous laboratory studies show that individuals avoid information concerning the potentially adverse welfare consequences of their self-interested decisions (Ehrich and Irwin, 2005; Dana, Weber, and Kuang, 2007; Grossman and Van der Weele, 2017; Serra-Garcia and Szech, 2022).³ Two studies contribute related field evidence: Andreoni, Rao, and Trachtman (2017) illustrate information avoidance in the realm of charitable giving, whereas Mechtenberg et al. (2024) investigate self-signaling and its impact on information processing in the context of moral voting.

My research adds to this literature in several ways. To give rise to what is referred to as “moral wiggle room” (Dana, Weber, and Kuang, 2007), some degree of uncertainty must be present in the moral decision-making situation. In the existing motivated Bayesian literature, the uncertainty stems from the question of whether or not one’s self-serving behavior could have adverse welfare consequences and, therefore, would violate the norm “do not harm others.” In my setting, the question as to whether a social norm is applicable is uncontroversial in that the subjects know that certain self-serving behaviors violate a social norm. In my experiment, uncertainty arises from the participant’s assessment of whether others conform to this norm. My contribution lies in investigating whether individuals will seek yet another source of uncertainty to rationalize their actions. Arguably, this constitutes a more intricate approach to belief manipulation, which arises once the straightforward opportunity to distort one’s beliefs has been eliminated. Moreover, within this literature, it is common to investigate belief distortion as a phenomenon that either precedes or occurs concurrently with a given behavior. In my scenario, participants distort their beliefs to rationalize their past actions.

The relevant social norm in my experimental set-up is the one of behaving honestly. Lying is of economic relevance in numerous situations, such as tax evasion, corporate fraud, or asymmetric information in principal-agent relationships.⁴ Existing studies have

²For an overview of the literature on information avoidance in various contexts see Golman, Hagmann, and Loewenstein (2017). In moral settings, the term “willful ignorance” has been coined to refer to such behavior (Ehrich and Irwin, 2005; Grossman and Van der Weele, 2017).

³For a literature review prior to 2016, cf. Gino, Norton, and Weber (2016).

⁴One can take different stances on the question of whether lying is immoral per se. I consider the setup constructed in my work as a “moral setting.” However, one could disagree and choose to view the

established that people include moral considerations in their decision-making and resist being dishonest (e.g., Fehr and Schmidt, 2006; Abeler, Nosenzo, and Raymond, 2019). Economic approaches often assume (heterogeneous) internal lying costs (Gibson, Tanner, and Wagner, 2013; Abeler, Becker, and Falk, 2014). A study by Bicchieri, Dimant, and Sonderegger (2023) which explores whether subjects manipulate their beliefs regarding the compliance of others with an honesty norm, is most closely related to my work. Their research demonstrates that people convince themselves that lying behavior undertaken to exploit an upcoming lying opportunity for personal gain, is widespread.

My work differs from the study by Bicchieri, Dimant, and Sonderegger (2023) in several ways. First, and most importantly, these authors explore how belief formation is shaped by anticipation of a future decision. My work asks whether the bar for manipulating one's beliefs is even lower. I investigate whether subjects distort their beliefs not to pave the way for future personal gain but solely to view their own past behavior in a positive light.⁵ Second, I introduce two different temptation environments, one of which makes lying more appealing. This allows for evidence that lying is context-specific and for investigating the impact of institutions on beliefs. My research highlights how inadequately designed institutions not only lead to dishonest behavior but also negatively affect individuals' perceptions of others. This underscores the spectrum of detrimental effects associated with poorly designed institutions.

Belief manipulation concerning the behavior of others has also been studied by Di Tella et al. (2015) and Ging-Jehli, Schneider, and Weber (2020), although they focus on different aspects than the norm compliance of peers. Similarly to Bicchieri, Dimant, and Sonderegger (2023), these two studies examine instrumental motives of belief manipulation, while my aim is to explore hedonic motives. One study explicitly addressing hedonic belief distortion is Galeotti, Saucet, and Villeval (2020). However, rather than investigating beliefs about others' norm compliance, they are interested in beliefs individuals have

results of my work in terms of "breaking a norm," while not behaving immorally. For further discussion, cf. appendix.

⁵In the literature, the former cases have been referred to as *instrumental* or *strategic* belief manipulation while the latter case has been called *hedonic* moral cleansing (Galeotti, Saucet, and Villeval, 2020).

about their own past behavior.⁶

To examine such manipulation of beliefs regarding norm compliance of peers for hedonic reasons, I develop a novel experimental approach to show that behaving immorally negatively impacts what people think others would do in the same situation. I use a large-scale incentivized online experiment with a representative sample of the general UK population. To exogenously vary behavior in the two groups, I expose subjects to one of two situations, in which lying is more tempting in the first than in the second. I find that my treatment increases the proportion of people who choose to lie. Subsequently, lying negatively affects beliefs about other people. In all treatments, subjects are overly pessimistic concerning the behavior of others. I include two additional treatments to distinguish the possible reasons why beliefs vary between the two situations. This approach shows that some differences in beliefs are caused by belief manipulation.

The remainder of this paper is structured as follows: Section 2 develops the main hypotheses. Section 3 introduces my experimental design and procedures. The empirical results are presented in section 4. Section 5 provides an additional discussion of the results. Section 6 gives the conclusion.

3.2 Hypotheses

I design the two temptation environments such that lying is more tempting in one scenario than in the other. The two environments differ in terms of three components. First, the task the participants perform varies slightly (die-roll task vs. mind game, Kajackaite and Gneezy, 2017). Second, I provide higher monetary incentives for lying in the tempting environment. Third, I make it very salient that lying is lucrative and that there is no risk of being caught (for more details, cf. the section on the experimental design). I expect these three components to lead to a higher share of participants lying in the group exposed

⁶Advancing the understanding of how humans form their perceptions of other people is particularly important, as it lays the foundation for trust, which, in turn, plays a pivotal role in a society's functioning and a country's economic success (La Porta et al., 1997; Knack and Keefer, 1997; Zak and Knack, 2001; Bloom, Sadun, and Van Reenen, 2012).

to a higher temptation. This effect is a precondition for testing my main hypothesis H_2 .

H₁: More people lie if the temptation is high than if the temptation is low (i.e., lying is context-specific).

The literature in psychology has a long tradition that shows that people exert a great deal of effort to maintain a positive self-image. Self-deception and motivated cognition are important tools to achieve this goal (e.g., Steele, 1988; Kunda, 1990). More recently, economists have begun to investigate motivated beliefs in economic contexts (e.g., Dana, Weber, and Kuang, 2007; Bénabou and Tirole, 2011; Gino, Norton, and Weber, 2016).⁷ In my scenario, the subjects who have lied will be compelled to manipulate their beliefs about what others would do in this situation, thus, about how common norm violation is.⁸ Importantly, I am interested in belief manipulation due to having lied. I, therefore, control for the effect of the temptation environment on beliefs (cf. section 3).

H₂ (Main hypothesis): Lying leads to more pessimistic beliefs about others. People exposed to a high temptation to lie have more pessimistic beliefs about others who are in the same situation (after controlling for the effect of the temptation environment).

3.3 Experimental Design

3.3.1 Experimental Task and Main Outcome Variable

The central component of the experiment in this study is the die-roll task (Fischbacher and Föllmi-Heusi, 2013). Each participant rolls a six-sided die and reports the number.

⁷In the existing literature subjects have some wiggle room as to whether the norm applies in their particular case. My contribution is to analyze a setting in which the norm applies uncontroversially, but subjects can still manipulate their beliefs about whether others would adhere to the norm in the same situation.

⁸In general there are other strategies to preserve a positive self-image, such as shifting blame to others (Bartling and Fischbacher, 2012; Oexl and Grossman, 2013), or forgetting one's own past actions (Galeotti, Saucet, and Villeval, 2020). Note that subjects who bias their beliefs forgo monetary rewards as the belief elicitation is incentivized in my experiment.

One winning number results in a bonus, and the other five outcomes of the die-roll yield no payoff. The experimenter does not observe which number a participant has rolled. Therefore, there is no way for them to determine whether an individual has told the truth when reporting their die-roll outcome.⁹ On an aggregate level, one can predict that 1/6 of all participants see the winning number; thus without any lying 16.67% of all subjects should report the winning number. Therefore, the share of participants lying in each group can be determined.¹⁰

The main outcome variable is called “beliefs about others.” I ask subjects the following incentivized question (*after* their own reporting of the number): “What do you think: What was the percentage of people in a previous study who have reported their rolled number to be equal to their winning number?” Besides the question, the study provides additional explanations to ensure that participants understand the statistical intuition, e.g., the fact that answering 16.67% would indicate that no one has lied (cf. appendix for specific wording).

3.3.2 Experimental Dimension I: High vs. Low Temptation to Lie

The usual gold standard in experimental science is to randomize participants at an individual level into treatment groups. In this context, this would entail randomizing subjects into “lying” and “truth-telling.” However, such treatment is not feasible as this study is concerned with human behavior that was chosen in a self-determined manner.¹¹ I, therefore, design two decision-making environments to make participants more/less likely to

⁹One potential concern is that participants fear deception and suspect being secretly monitored despite being promised otherwise. To increase participants’ trust in not being monitored, I offer them the possibility of rolling a physical die and provide a link to an external website where they can generate a die-roll. While monitoring an external website is possible in theory, it would be technically very complicated as soon as subjects leave the website of the experimental interface.

¹⁰This paradigm has been extensively employed in the literature. In addition, some studies have illustrated that dishonest misreporting in this task is associated with morally questionable actions outside the lab (Cohn, Maréchal, and Noll, 2015; Potters and Stoop, 2016; Hanna and Wang, 2017; Cohn and Maréchal, 2018; Dai, Galeotti, and Villeval, 2018).

¹¹If one dictated to the participants to lie in one condition, lying might not have any effect on participants’ beliefs about others because they would not perceive to have violated a norm.

lie. I call these two environments HIGH and LOW. Generally speaking, simply increasing the monetary incentives for dishonesty would not lead to a higher percentage of people choosing to lie (Abeler, Nosenzo, and Raymond, 2019). Based on results from existing literature, I vary three components across the two environments.

The first and most important component is a slight variation in the task. While participants under LOW play the standard die-roll task as described above (Fischbacher and Föllmi-Heusi, 2013), subjects under HIGH play the mind game (Kajackaite and Gneezy, 2017). In the former, the number five is the winning number and in the latter, the subjects choose their own winning number.¹² Previous literature has shown that adjusting the die-roll task by implementing the mind game is an effective way to increase lying (Kajackaite and Gneezy, 2017).

The second component is a variation in the monetary incentives. The monetary reward under HIGH is ten times higher than the reward under LOW (£2.5 vs. £0.25). Existing literature has shown that, somewhat contrary to economic intuition, raising monetary incentives does not systematically lead to more lying in the standard die-roll game (Fischbacher and Föllmi-Heusi, 2013; Kajackaite and Gneezy, 2017; Abeler, Nosenzo, and Raymond, 2019). However, under the mind game, higher monetary incentives do cause more lying (Kajackaite and Gneezy, 2017).

The final component is an adjustment in the wording. The instructions in both HIGH and LOW inform participants that there would be a monetary reward for reporting the winning number (which in most cases means lying) and that there is no risk of getting caught out. However, this information is particularly emphasized under HIGH (cf. appendix for specific wording). Existing literature in economics (Kajackaite and Gneezy, 2017) as well as in social psychology (Vrij, 2008) argues that the fear of being caught out is an important determinant of lying behavior, therefore, my study's treatment component is designed to (further) decrease subjects' concerns about being exposed when lying.¹³

¹²More specifically, subjects under HIGH see an additional screen asking them to choose a number between one and six as their winning number and to memorize it before moving on to the next screen. This is a private choice and is not reported. For the specific wording, cf. the appendix.

¹³Importantly, I argue that while adjusting the wording stresses that there is no reason to be concerned

Importantly, I vary all three components simultaneously. Hence, I cannot disentangle the effects of the separate factors. However, quantifying separate effects is not the goal of this treatment dimension. All modifications aim to vary the share of people choosing to lie in these two environments.

3.3.3 Experimental Dimension II: Active vs. Passive

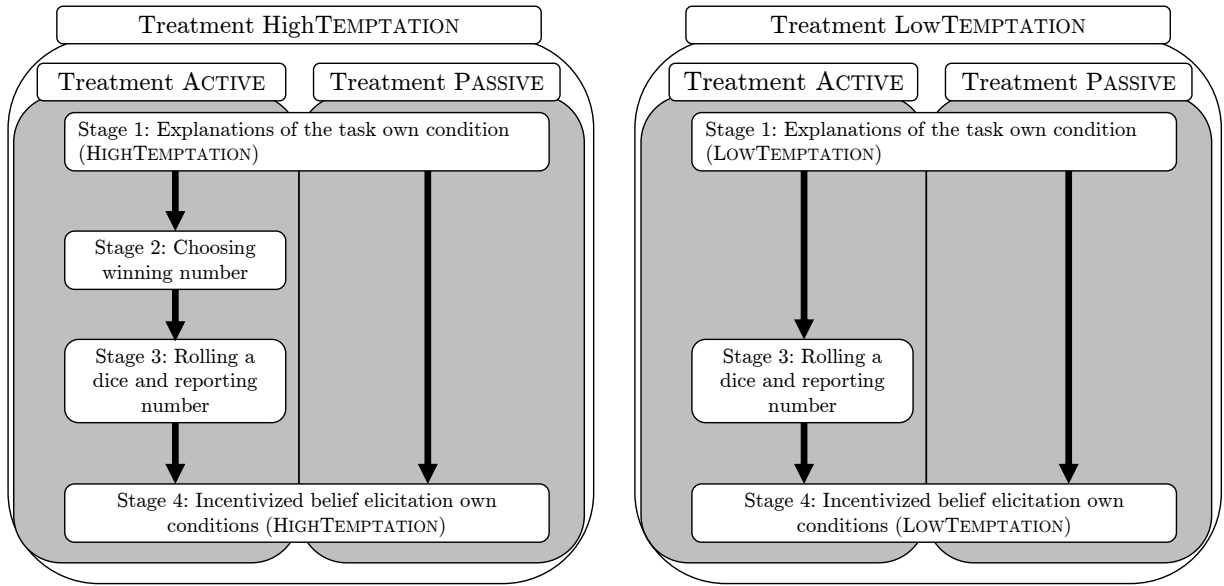
I elicit the main outcome variable “beliefs about others” with respect to each participant’s own condition. In other words, subjects state their beliefs about the behavior of other people under the same condition (either HIGH or LOW).¹⁴ The difference in beliefs between HIGH and LOW will partly occur because subjects (rationally) anticipate the effect of the different temptation levels on lying. I introduce the second treatment dimension to control for this. In the PASSIVE condition, participants are not allowed to roll any die. They are only asked to report their beliefs. These two conditions are identical, apart from the stage in which the die is rolled. (cf. Figure 3.1). Hence, any difference between ACTIVE and PASSIVE (when holding the other dimension constant) must be due to having been in the decision-making situation and having chosen one’s own behavior. This approach will be further discussed in the results section. In sum, the experiment includes four treatments (HIGHACTIVE, HIGHPASSIVE, LOWACTIVE, LOWPASSIVE) that vary on two dimensions (HIGH vs. LOW and ACTIVE vs. PASSIVE), as illustrated in Figure 3.1. The appendix provides some additional discussion on the interdependence between behavior and beliefs.

3.3.4 Payment Scheme and Experimental Procedure

I programmed and conducted the experiment using oTree (Chen, Schonger, and Wickens, 2016) and recruited participants via Prolific (Palan and Schitter, 2018). Data were col-

about being exposed as a liar, it does not legitimize lying and hence does not change the underlying norm relative to LOW.

¹⁴Using this procedure, the main outcome variable for active players refers to a situation they have experienced and not to a hypothetical situation. Importantly, it also avoids revealing the existence of another experimental condition to the subjects.

Figure 3.1 Visualization experimental design

lected between October 8, 2021, at 16:13, and October 10, 2021, at 22:38. I collected all the data in a single session. The sample is representative of the UK’s national population and has been stratified across three demographic dimensions: age, gender, and ethnicity. This work was registered with aspredicted.org and IRB approval (ethical clearance) was obtained at the University of Hamburg on September 28, 2021 (cf. appendix). The pre-registration states the required (final) sample size of 1200 and four exclusion criteria.¹⁵ On average, participants took 6.2 minutes to complete the study and they earned £1.38. This is equivalent to an average hourly wage of more than £13.¹⁶

3.4 Results

Overall, 1201 subjects are included in the analysis. Table 1 presents descriptive statistics for the sample. Subjects are between 18 and 88 years old, averaging 44.86 years. Of the participants, 49% are female. English is the first language for most (89%) participants

¹⁵A participant was excluded from the analysis in the following cases: 1.) They did not complete the whole study. 2.) They completed the experiment in an unrealistically short period of time: less than 01:20. 3.) They took an unrealistically long time to finish the experiment: more than 5 standard deviations longer than the mean. 4.) They provided an unreasonable answer to the question regarding their beliefs about others (i.e., that less than 1/6 reported the winning number).

¹⁶The median hourly wage in the UK in 2021 was £13.57 (UK-Government, 2022).

and the average prolific score is over 99 (out of 100). Of all active players, 38% report having rolled the winning number and, on average, participants believe that 53.94% of the other players in the same situation would report the winning number. These two main outcome variables are also analyzed for each treatment in the following.

Table 3.1 Summary statistics

	Mean	S.d.	Min	Max	N
Age	44.86	15.67	18	88	1200
Gender	0.49	0.50	0	1	1201
Prolific score	99.41	1.47	77	100	1201
First language	0.89	0.32	0	1	1201
Share winning number	0.38	0.48	0	1	606
Beliefs about others	53.94	24.48	17	100	1201

Notes: Gender = 1 if male and 0 otherwise. First language = 1 if English and 0 otherwise.

Table 3.2 Treatment effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full sample	HA	HP	LA	LP	p-value (MWU) HA - LA	p-value (MWU) HP - LP
Share winning number	0.38 (0.485)	0.44 (0.497)	-	0.31 (0.464)	-	0.0012**	-
Beliefs about others	53.94 (24.48)	58.28 (24.94)	52.27 (23.81)	51.59 (24.17)	53.63 (24.55)	0.0010**	0.4872
N	1201	302	300	304	295		

Notes: Mean coefficients. Standard deviations in parentheses. Abbreviations: HA = HIGHACTIVE, HP = HIGHPASSIVE, LA = LOWACTIVE, LP = LOWPASSIVE. “Share of winning number” refers to the share of subjects who report the winning number. I exclude participants from the analysis if their beliefs about others is lower than 16.67. The instructions clearly outline the logic of the task indicating that a guess below 16.67 could not be meaningful (as preregistered). * indicates significance at the 10% level, ** at the 5% level, and *** at the 1% level.

The results of the treatment effects are presented in Table 3.2. One can see that under HIGH more people report the winning number than under LOW (13 percentage points difference: 44% - 31%, significant on a 5%-level, also cf. Figure 3.2). In other words, lying

appears to be context-specific and this treatment dimension is successful in influencing the share of people who lie. This is a prerequisite for investigating H_2 .

Result 1: More people lie when the temptation is high than when the temptation is low (i.e., lying is context-specific).

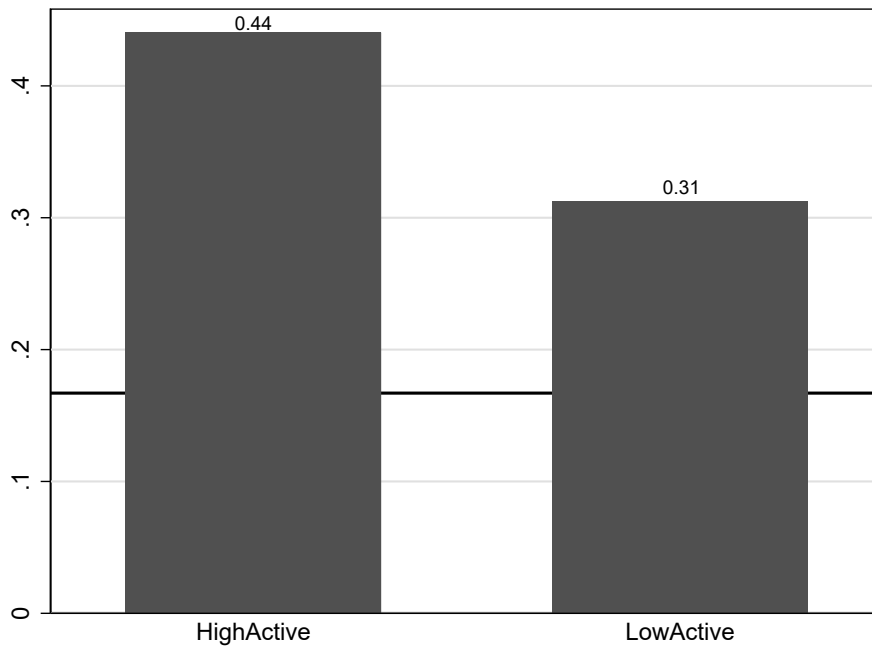
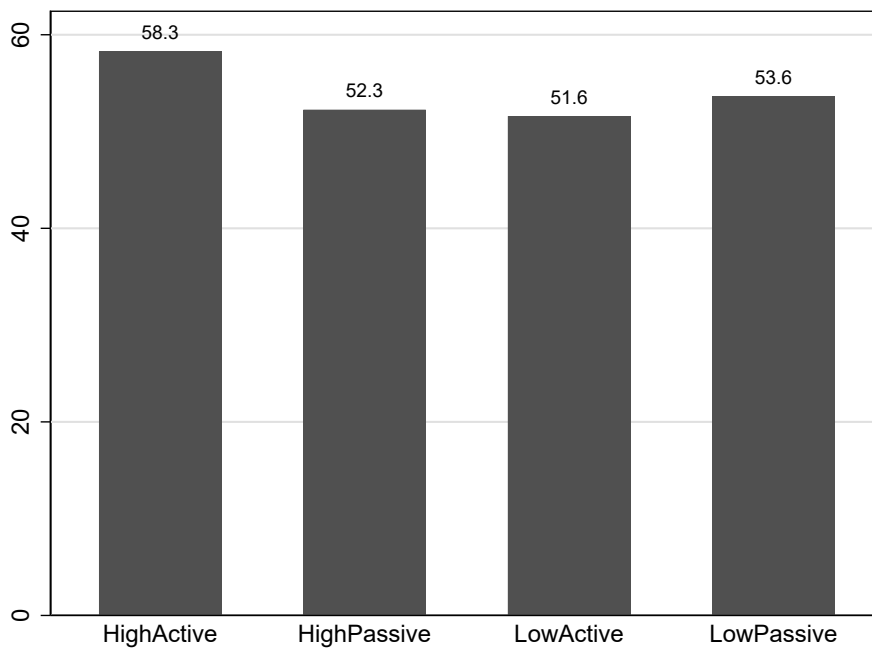
One can now assess the effect of the temptation environment on beliefs. There are two notable aspects of these results (cf. Table 3.2 and Figure 3.3). First, compared with actual behavior, beliefs about others' behavior are widely inaccurate and overly pessimistic. While 31% (44%) report the winning number under LOW (HIGH), participants in any treatment, on average, believe that over 50% report having rolled the winning number. Second, there is a significant difference in beliefs between HIGH and LOW for active players (HIGHACTIVE: 58.82%, LOWACTIVE : 51.59%). Hence, people have more pessimistic beliefs about others when exposed to a situation in which lying is more tempting. When we turn to passive players, the difference between HIGH and LOW is small and insignificant (HIGHPASSIVE: 52.27%, LOWPASSIVE: 53.63%).¹⁷

Importantly, comparing beliefs under HIGHACTIVE and LOWACTIVE does not sufficiently address H_2 , because those beliefs evaluate behavior under different circumstances. The difference in beliefs about others between HIGHACTIVE and LOWACTIVE might (partly) occur because subjects (rationally) anticipate the varying effect of the temptation environments on lying between the two treatments. To put it differently:

$|HIGHACTIVE - LOWACTIVE|$: *anticipated effect of temptation environment + effect of own behavior (i.e., lying or truth-telling)*

$|HIGHPASSIVE - LOWPASSIVE|$: *anticipated effect of temptation environment*

¹⁷One could also look at the results from a different perspective: Comparing HIGHPASSIVE and HIGHACTIVE, one can see that beliefs are more pessimistic in the latter (HIGHPASSIVE: 52.27%, HIGHACTIVE: 58.28%). The average belief under LOWPASSIVE (53.63%) is more pessimistic than under LOWACTIVE (51.59%), but the difference is not statistically significant. This perspective is not the focus of my work, but the latter result is noteworthy and will be discussed in section 5. Further, it is noteworthy that there is no evidence that active players deliver more accurate predictions (due to having experienced the situation). This nuance will also be discussed in section 5.

Figure 3.2 Share of subjects reporting winning number**Figure 3.3** Average beliefs about others

The next step is therefore to analyze the effect $|\text{HIGHACTIVE} - \text{LOWACTIVE}|$ while controlling for $|\text{HIGHPASSIVE} - \text{LOWPASSIVE}|$. In other words, this study is interested in the difference between two differences: the effect of having lied while controlling for the effect of the temptation level of the environment. To achieve this, I estimate the following equation:

$$B_i = \beta_0 + \beta_1 T_i^H + \beta_2 T_i^A + \beta_3 (T_i^H \cdot T_i^A) + \Theta X_i + u_i \quad (3.1)$$

where B_i represents the outcome variable, i.e., beliefs about others. T_i^H is a dummy that takes the values one if individual i is randomly exposed to the HIGH treatment dimension and zero otherwise. T_i^A is a dummy that takes the values one if individual i is randomly exposed to the ACTIVE treatment dimension and zero otherwise. $\hat{\beta}_3$ is the difference-in-difference estimate. X_i is the set of control variables including controls for age, gender, prolific score, and first language.¹⁸

The regression output table shows that the coefficient of interest is significant at the 5%-level and robust to including controls in the regression specification. In other words, people in the more tempting environment have more pessimistic beliefs even when controlling for the temptation level of the two environments, providing evidence that lying negatively impacts the individual's beliefs about others in the same situation.¹⁹

Result 2: Lying leads to more pessimistic beliefs about how others would behave. The group exposed to a higher temptation to lie has more pessimistic beliefs about others' possible behavior in the same situation (after controlling for the effect of the temptation environment).

¹⁸Notably one could argue that this approach does not fully control for the temptation environment, since the passive players cannot accurately assess the temptation without having been in that situation. This point will be addressed in section 5.

¹⁹While the interaction effect is significant, the two main effects are not. Note that the table presents *average* effects of HIGH and ACTIVE. It is not surprising that these average effects are insignificant since the effect of HIGH goes in opposite directions for ACTIVE and PASSIVE and the effect of ACTIVE goes in opposite directions for HIGH and LOW (cf. Table 3.2). The central information of this output table is the interaction coefficient; the effects of HIGH and ACTIVE are better examined separately for each condition.

Table 3.3 OLS regression results

	(1)	(2)
	Beliefs about others	Beliefs about others
HIGH	-1.361 (1.986)	-1.485 (1.944)
ACTIVE	-2.045 (1.994)	-2.233 (1.962)
HIGH x ACTIVE	8.050** (2.818)	7.940** (2.761)
Controls	NO	YES
Constant	53.63*** (1.432)	9.249 (63.69)
N	1201	1200
R^2	0.011	0.060

Notes: Robust standard errors (HC3, Long and Ervin, 2000) in parentheses. Controls include age, gender, prolific score, and first language. HIGH = 1 if the treatment is HIGH. ACTIVE = 1 if the treatment is ACTIVE. * indicates significance at the 10% level, ** at the 5% level, and *** at the 1% level.

This result is in line with Bicchieri, Dimant, and Sonderegger (2023), who find that people manipulate their beliefs about others, but focus on instrumental motives. This finding also relates to Galeotti, Saucet, and Villeval (2020) who examine beliefs about *one's own past behavior* finding that people manipulate their beliefs for instrumental reasons, while hedonic reasons do not provide sufficient explanations.

3.5 Discussion

3.5.1 Alternative Mechanisms: False Consensus Effect

My data suggest that beliefs about others are influenced by an individual's own behavior. While this paper refrains from making strong claims regarding the underlying mechanisms, I have argued that the results are driven by those subjects who have behaved dishonestly and subsequently form motivated beliefs. One potential alternative explanation is the

“false consensus effect” (Ross, Greene, and House, 1977). According to the bias, *all* people engaging in active behavior (not only ones breaking the norm) tend to form biased beliefs. Mullen et al. (1985) summarize this bias as follows: “People who engage in a given behavior will estimate that behavior to be more common than it is estimated to be by people who engage in alternative behaviors” (also, cf. Offerman, Sonnemans, and Schram, 1996; Engelmann and Strobel, 2000, 2012).

Applied to my setting, the predictions of the “false consensus effect” and “motivated beliefs” differ in a critical way. According to the former, everyone who has made an active decision will form biased beliefs, while according to the latter, only people who have acted dishonestly will be inclined to manipulate their beliefs. From this perspective, the data are in line with motivated beliefs, rather than with the consensus effect. Examining Table 3.2, one can compare active and passive players, that is, compare people who face having to take a decision with people who are passive bystanders. Here, one can see that active players under HIGH have significantly more pessimistic beliefs than the corresponding passive bystanders (MWU-test p-value = 0.0024). Looking at LOW, there is no significant difference between active and passive players (MWU-test p-value = 0.3401). This is consistent with the concept of “motivated beliefs,” as it suggests that differences are driven by the beliefs of those who have lied, rather than by the distorted beliefs of everyone who has been active.

3.5.2 Alternative Mechanisms: Own Behavior vs. Behavior of a Random Participant

Considering one’s own behavior when forming beliefs about others can be sensible as Engelmann and Strobel (2000) point out. They emphasize that it is perfectly rational to include one’s own behavior in forming beliefs about a certain group if one is a member of the group in question. They contend that such behavior is only irrational if individuals put more weight on their behavior than on the behavior of a randomly chosen person from the group. Further, their experimental evidence contradicts the notion that people

overrate their own behavior.

This subsection investigates whether my results could be driven by a mechanism in the spirit of Engelmann and Strobel (2000) rather than by strategic belief formation as I have proposed. Applying the idea of Engelmann and Strobel (2000) to my work would imply that being active (rather than passive) provides subjects with information about the behavior of one individual (i.e., themselves) from the population whose behavior they are being asked to estimate. If this mechanism was a driving factor, the effect of the ACTIVE dimension would be similar to providing subjects with information about the behavior of one random other participant. Consequently, I design an additional treatment, HIGHPASSIVEINFOOTHER, that is almost identical to HIGHPASSIVE, but includes an additional sentence, which reads: “We will provide you with the information about the behavior of one other participant: The first participant from a previous study has reported the number 4, the number which gave her a bonus.”²⁰

My data show that while the average of the HIGHPASSIVE treatment is 52.27 (23.81 s.d.), the average of HIGHPASSIVEINFOOTHER is 52.22 (25.26 s.d.). The p-value of the MWU-test is 0.9285. Hence, including the information about the behavior of another random subjects has no effect. In other words, the difference between HIGHPASSIVE and HIGHACTIVE is not likely to be driven by an effect in the spirit of Engelmann and Strobel (2000). The effect I observe must be due to another reason, such as the strategic manipulation of one’s own beliefs.

3.5.3 Caveats, Parallel Trend Assumption, and Lying as an Experience Good

One potential caveat of the empirical approach is the specification of the difference-in-difference analysis. One underlying assumption of this approach is that the “anticipated effect of the temptation environment” (cf. result section) must be the same for the active

²⁰This is an exploratory additional treatment that has not been preregistered. This additional treatment has 144 observations.

players as well as for the passive players.²¹ One could argue that active players perceive the effect differently because they experience the decision-making situation themselves. In other words, lying would be an experience good, and one can only assess the costs of lying after encountering the situation. That would imply that participants would deliver better predictions of others' behavior in the active conditions. However, there is no clear evidence of this in the data. In the LOW condition, active subjects are closer to the true answer than passive subjects (but the difference is statistically insignificant); the opposite is the case under the HIGH condition. Beliefs about others are 52.27% under HIGHPASSIVE and 58.28% under HIGHACTIVE (the true value being 44%). Hence, under HIGH, the active players are, on average, approximately 6 percentage points less accurate than their passive counterparts. This suggests that being in the decision-making situation does not give subjects a better capacity to assess the behavior of others.

Overall, I am cautious in interpreting the size of the point estimate based on the difference-in-difference estimate. However, even if one considers the exact effect size to be unclear and has doubts regarding the parallel trend assumption, the raw values of the main outcome variable "beliefs about others" (cf. Table 3.2) show a clear picture. Beliefs about others' likely behavior under HIGHACTIVE are much more pessimistic than in any other condition, and it seems evident that lying has some effect on these beliefs. In the same vein, even if one is convinced that the false consensus effect or other underlying mechanisms are at play, it is hard to ignore that beliefs under HIGHACTIVE stand out, which suggests that motivated beliefs are a driving mechanism.

Furthermore, one might find it striking that there is no significant difference between the average belief under LOWACTIVE and LOWPASSIVE (cf. Table 3.2) and might also consider this finding to be at odds with H_2 . According to the hypothesis, lying leads to more pessimistic beliefs and since there are no participants who have lied in LOWPASSIVE and some people who have lied in LOWACTIVE, one could expect to see a difference in average beliefs. One explanation for why we do not see a difference is the small share of people

²¹When applied as an econometric tool for observational data to mimic an experimental setting, this assumption is commonly called the "parallel trend assumption," and most often refers to a time component.

who lie in the active condition. Only roughly 14.33% (31% – 16.67%) of participants lie under LOWACTIVE and this share might not be sufficiently high to significantly alter the group’s overall average. Also, note that the preregistered specification of this work for testing H_2 is the difference-in-difference approach discussed in the result section, rather than comparing the averages of HIGHACTIVE with HIGHPASSIVE and LOWACTIVE with LOWPASSIVE separately. Investigating the underlying mechanisms of this finding further might be an interesting direction for future research.

3.5.4 Non-causal Evidence That Lying Causes More Negative Beliefs

The questionnaire I administer after the experiment contains additional evidence suggesting that the results are indeed driven by subjects who choose to lie and subsequently engage in strategic manipulation of beliefs. In the questionnaire, (active) participants are asked to voluntarily disclose whether they had lied during the experiment.²² Although such self-reported data should be viewed with caution, there is little reason to believe that subjects who have not lied will decide to report engaging in dishonest behavior. 26 subjects in the HIGHACTIVE condition and 26 subjects in the LOWACTIVE condition choose to report lying behavior. The average belief about others’ behavior among these 52 subjects is 78.5%. The beliefs in this group are tremendously more negative than the average belief in any of the four treatment groups (more than 20 percentage points higher than in HIGHACTIVE, the treatment with the most negative beliefs). While this observation is based on self-reported data and cannot be considered causal, it is consistent with the reasoning and analysis presented in this paper. Examining the data in this way shows a strong correlation between admitting to having lied and having negative beliefs. This suggests that having lied is an important driving factor of the main results.

²²The question has the following wording: “Your earnings have been determined and the following question has no consequences for either your earnings, the approval of your submission, or your rating on Prolific.co. As promised, the outcome of your die-roll is not known to us. We, therefore, don’t know if you have been truthful or not when reporting the outcome of your die-roll. For our research purposes, we would kindly ask you to voluntarily share this information by choosing one of the following options.”

3.6 Conclusion

People sometimes find themselves in situations where behaving dishonestly seems irresistible, even if they would prefer to have behaved honestly. This prompts the following question: Once people give in to the temptation of lying, how does lying change the way they see other people? I designed an incentivized experiment and conducted an online study with a representative sample of the general UK population. My work provides evidence that lying causes pessimistic beliefs regarding the behavior of others. In my setting, I argue that belief manipulation occurs when individuals seek to maintain a positive self-image.

The insight that lying causes negative beliefs raises the following question for future research: Do people strategically adjust their *reported* beliefs after learning this evidence? Being informed about this relationship between stating negative beliefs about others and having lied oneself, subjects might strategically present their own beliefs as optimistic so that they will be considered honest.

Broadly, these findings show that deficient institutions could be even more harmful than is commonly understood. In my study, they did not only make people behave dishonestly; in addition, they subsequently caused more pessimistic beliefs about the behavior of others. This is particularly alarming, considering the evidence that humans tend to attribute perceived misbehavior to personal characteristics rather than to factors in the environment (Han, Liu, and Loewenstein, 2023). A fruitful direction for future research would be to examine the scope of these negative impressions of others and whether they also translate into pessimistic beliefs regarding behavior in other situations. Do they, e.g., decrease trust and hamper cooperation? These questions are important for a better understanding of the interplay between beliefs and behavior. I consider the current findings an important step toward finding such answers.

Acknowledgments

Valuable discussions with the following people have greatly improved this paper: Roland Bénabou, Robert Dur, Marc Kaufmann, Balázs Kruspe, Andreas Lange, Lydia Mechtenberg, Elise Payzan-Le Nestour, Daniele Nosenzo, Timo Promann, Felix Schafmeister, Silvia Sonderegger, and Johannes Walter. I also want to thank the participants of AS-FEE 2021, NYU CESS 2021, ASFEE 2022, and the UHH Collective Decision-Making PhD seminar 2022. I am grateful to David Lucius and Ziad Zorgati for outstanding technical assistance, Timo Promann for important technical advice, and Christine Anthonissen for proofreading. This research project is funded by the Deutsche Forschungsgemeinschaft (DFG) - GRK 2503.

Bibliography

- Abeler, Johannes, Anke Becker, and Armin Falk. 2014. “Representative evidence on lying costs.” *Journal of Public Economics* 113:96–104.
- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. “Preferences for truth-telling.” *Econometrica* 87 (4):1115–1153.
- Allen IV, James, Arlete Mahumane, James Riddell IV, Tanya Rosenblat, Dean Yang, and Hang Yu. 2021. “Correcting perceived social distancing norms to combat COVID-19.” *NBER Working Paper* .
- Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2017. “Avoiding the ask: A field experiment on altruism, empathy, and charitable giving.” *Journal of Political Economy* 125 (3):625–653.
- Bartling, Björn and Urs Fischbacher. 2012. “Shifting the blame: On delegation and responsibility.” *The Review of Economic Studies* 79 (1):67–87.
- Bénabou, Roland and Jean Tirole. 2006. “Incentives and prosocial behavior.” *American Economic Review* 96 (5):1652–1678.
- . 2011. “Identity, morals, and taboos: Beliefs as assets.” *The Quarterly Journal of Economics* 126 (2):805–855.
- . 2016. “Mindful economics: The production, consumption, and value of beliefs.” *Journal of Economic Perspectives* 30 (3):141–164.
- Bicchieri, Cristina. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.
- Bicchieri, Cristina, Eugen Dimant, and Silvia Sonderegger. 2023. “It’s not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion.” *Games and Economic Behavior* 138:321–354.

- Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2012. “The organization of firms across countries.” *The Quarterly Journal of Economics* 127 (4):1663–1705.
- Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. “From extreme to mainstream: The erosion of social norms.” *American Economic Review* 110 (11):3522–3548.
- 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.
- Cohn, Alain and Michel A. Maréchal. 2018. “Laboratory measure of cheating predicts school misconduct.” *The Economic Journal* 128 (615):2743–2754.
- Cohn, Alain, Michel A. Maréchal, and Thomas Noll. 2015. “Bad boys: How criminal identity salience affects rule violation.” *The Review of Economic Studies* 82 (4):1289–1308.
- D’Adda, Giovanna, Michalis Drouvelis, and Daniele Nosenzo. 2016. “Norm elicitation in within-subject designs: Testing for order effects.” *Journal of Behavioral and Experimental Economics* 62:1–7.
- Dai, Zhixin, Fabio Galeotti, and Marie C. Villeval. 2018. “Cheating in the lab predicts fraud in the field: An experiment in public transportation.” *Management Science* 64 (3):1081–1100.
- Dana, Jason, Roberto A. Weber, and Jason X. Kuang. 2007. “Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness.” *Economic Theory* 33:67–80.
- 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):3416–3442.
- Ehrich, Kristine R. and Julie R. Irwin. 2005. “Willful ignorance in the request for product attribute information.” *Journal of Marketing Research* 42 (3):266–277.

- Engelmann, Dirk and Martin Strobel. 2000. “The false consensus effect disappears if representative information and monetary incentives are given.” *Experimental Economics* 3:241–260.
- . 2012. “Deconstruction and reconstruction of an anomaly.” *Games and Economic Behavior* 76 (2):678–689.
- Fehr, Ernst and Klaus M. Schmidt. 2006. “The economics of fairness, reciprocity and altruism – experimental evidence and new theories.” In *Handbook of the Economics of Giving, Altruism and Reciprocity*. 615–691.
- Fischbacher, Urs and Franziska Föllmi-Heusi. 2013. “Lies in disguise – an experimental study on cheating.” *Journal of the European Economic Association* 11 (3):525–547.
- Galeotti, Fabio, Charlotte Saucet, and Marie C. Villeval. 2020. “Unethical amnesia responds more to instrumental than to hedonic motives.” *Proceedings of the National Academy of Sciences* 117 (41):25423–25428.
- Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner. 2013. “Preferences for truthfulness: Heterogeneity among and within individuals.” *American Economic Review* 103 (1):532–548.
- Ging-Jehli, Nadja R., Florian H. Schneider, and Roberto A. Weber. 2020. “On self-serving strategic beliefs.” *Games and Economic Behavior* 122:341–353.
- 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.
- Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. “Bribing the self.” *Games and Economic Behavior* 120:311–324.
- Golman, Russell, David Hagmann, and George Loewenstein. 2017. “Information avoidance.” *Journal of Economic Literature* 55 (1):96–135.

- 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.
- Han, Yi, Yiming Liu, and George Loewenstein. 2023. “Confusing context with character: Correspondence bias in economic interactions.” *Management Science* 69 (2):1070–1091.
- Hanna, Rema and Shing-Yi Wang. 2017. “Dishonesty and selection into public service: Evidence from India.” *American Economic Journal: Economic Policy* 9 (3):262–290.
- Isenberg, Arnold. 1964. “Deontology and the ethics of lying.” *Philosophy and Phenomenological Research* 24 (4):463–480.
- Kajackaite, Agne and Uri Gneezy. 2017. “Incentives and cheating.” *Games and Economic Behavior* 102:433–444.
- Knack, Stephen and Philip Keefer. 1997. “Does social capital have an economic payoff? A cross-country investigation.” *The Quarterly Journal of Economics* 112 (4):1251–1288.
- Konow, James. 2000. “Fair shares: Accountability and cognitive dissonance in allocation decisions.” *American Economic Review* 90 (4):1072–1092.
- Kunda, Ziva. 1990. “The case for motivated reasoning.” *Psychological Bulletin* 108 (3):480–498.
- La Porta, Rafael, Florencio Lopez-de Silanes, Andrei Shleifer, and Robert W. Vishny. 1997. “Trust in large organizations.” *American Economic Review* 87 (2):333–338.
- Long, J. Scott and Laurie H. Ervin. 2000. “Using heteroscedasticity consistent standard errors in the linear regression model.” *The American Statistician* 54 (3):217–224.
- Maggian, Valeria. 2019. “Negative externalities of cheating: An experiment with charities.” In *Dishonesty in Behavioral Economics*. 183–191.
- Mechtenberg, Lydia, Grischa Perino, Nicolas Treich, Jean-Robert Tyran, and Stephanie W. Wang. 2024. “Self-signaling in voting.” *Journal of Public Economics* Forthcoming.

- Mullen, Brian, Jennifer L. Atkins, Debbie S. Champion, Cecelia Edwards, Dana Hardy, John E. Story, and Mary Vanderklok. 1985. “The false consensus effect: A meta-analysis of 115 hypothesis tests.” *Journal of Experimental Social Psychology* 21 (3):262–283.
- Oexl, Regine and Zachary J. Grossman. 2013. “Shifting the blame to a powerless intermediary.” *Experimental Economics* 16:306–312.
- Offerman, Theo, Joep Sonnemans, and Arthur Schram. 1996. “Value orientations, expectations and voluntary contributions in public goods.” *The Economic Journal* 106 (437):817–845.
- Palan, Stefan and Christian Schitter. 2018. “Prolific.ac – A subject pool for online experiments.” *Journal of Behavioral and Experimental Finance* 17:22–27.
- Potters, Jan and Jan Stoop. 2016. “Do cheaters in the lab also cheat in the field?” *European Economic Review* 87:26–33.
- Ross, Lee, David Greene, and Pamela House. 1977. “The “false consensus effect”: An egocentric bias in social perception and attribution processes.” *Journal of Experimental Social Psychology* 13 (3):279–301.
- Serra-Garcia, Marta and Nora Szech. 2022. “The (in)elasticity of moral ignorance.” *Management Science* 68 (7):4815–4834.
- Sliwka, Dirk. 2007. “Trust as a signal of a social norm and the hidden costs of incentive schemes.” *American Economic Review* 97 (3):999–1012.
- Sobel, Joel. 2020. “Lying and deception in games.” *Journal of Political Economy* 128 (3):907–947.
- Steele, Claude M. 1988. “The psychology of self-affirmation: Sustaining the integrity of the self.” In *Advances in Experimental Social Psychology*. 261–302.
- UK-Government. 2022. “Average hourly pay.” URL <https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/pay-and-income/average-hourly-pay/latest>.

Vrij, Aldert. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Zak, Paul J. and Stephen Knack. 2001. "Trust and growth." *The Economic Journal* 111 (470):295–321.

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." *American Economic Review* 110 (2):337–363.

Appendix A

Is Misreporting in the Die-Roll Task “Immoral”?

Existing literature has presented overwhelming evidence that people exhibit a preference for not lying²³ using the Fischbacher-Höllmi-Heusi paradigm (Abeler, Nosenzo, and Raymond, 2019). However, there may be differing opinions on whether dishonest behavior in this context can be considered “unethical” or “immoral”. In particular, if one does not view reporting untruths as unethical based on deontological reasons, the ethical judgment of such behavior relies on its consequences (“consequentialism”). One might argue that the harm caused by not telling the truth creates a burden on the budget of the experimenter, thus suggesting it is not morally wrong to choose an option that increases one’s own payoff at the expense of the experimenter’s remaining budget.

However, considering some empirical evidence, there are compelling reasons to view misreporting in this experimental context as unethical. First, Maggian (2019) conducted an experiment varying the negative externality of misreporting, either affecting the experimenter’s budget or donations to charity. Arguably, appropriating funds that one is not entitled to from charity (and thereby from disadvantaged individuals) constitutes immoral conduct. She reports no difference in misreporting between the treatment where misreporting hurts the budget of the experimenter and where misreporting diminishes donations made to charity. Fischbacher and Föllmi-Heusi (2013) report finding no effect when altering whether misreporting hurts the experimenter or another participant. Second, existing research has shown that dishonest misreporting in the lab is associated with morally questionable actions outside the lab, such as prisoners’ offenses against in-prison regulations (Cohn, Maréchal, and Noll, 2015), corrupt behavior by civil servants (Hanna and Wang, 2017), high school students engaging in improper behavior at school (Cohn and Maréchal, 2018) and passengers using public transport without paying for the ticket

²³In this work, I follow Isenberg’s definition of “lying”: “A lie is a statement made by one who does not believe it with the intention that someone else shall be led to believe it” (Isenberg, 1964, p. 466). For a definition of lying and deception in context which are commonly studied in economics see Sobel (2020).

(Dai, Galeotti, and Villeval, 2018). Further, Bicchieri (2005) argues that the vast majority of the behavioral literature on lying assumes that a truth-telling norm has been internalized.

Still, one might remain reluctant to view misreporting as unethical in this experimental task and emphasize that the outcome of a die-roll does not constitute a legitimate source for determining the distribution of pay-offs. In such a case, one might choose to view the results of my work in terms of “truth-telling” and “norm-following,” without adding a moral dimension when interpreting the results.

The Interdependence Between Behavior and Beliefs

On a general level, there are some gaps in the literature regarding the causal relationship between beliefs and behavior in moral dilemma situations: To what extent do beliefs impact behavior and to what extent does behavior influence beliefs? The stance that beliefs influence behavior is quite prominent in economics and several studies have provided evidence that exogenously treating beliefs leads to changes in behavior.²⁴

The reverse causal path is understudied and more complicated to analyze because behavior cannot be readily assigned exogenously. My experimental design provides an approach to address this challenge. The results of this work document how people distort their beliefs because of past behavior, even when there is no prospect of changing their behavior in the future. While I do not challenge the view that beliefs shape behavior, this study points out that beliefs are also shaped by (past) behavior and stresses a two-way interdependence between behavior and beliefs.

Since my research is designed to measure whether one’s behavior affects beliefs about others, I first elicit behavior, before measuring the beliefs. D’Adda, Drouvelis, and Nosenzo (2016) provide a discussion on how eliciting behavior first can have an effect on beliefs

²⁴A sizable body of literature has illustrated that social norms affect behavior (e.g., Sliwka, 2007; Bursztyn, Egorov, and Fiorin, 2020; Allen IV et al., 2021). Arguably, social norms are inevitably shaped by beliefs about others.

and how measuring beliefs first can have an impact on behavior.²⁵ Note that one could argue that regardless of when behavior and beliefs are measured, both of them might still emerge simultaneously at the point of time when the person decides how to behave. However, the underlying decision-making process is heterogeneous for different people. While some will immediately think about what others would do in the situation when confronted with the task at hand, others will not bother to reflect on the behavior of others until they are asked about it. Importantly, my experimental design does not make the question of “what would others do” salient at the point of time of the behavior and only asks this question later.

Appendix B

Screenshots of the experiment follow on the next pages.

²⁵The authors also conduct an experiment documenting little effect that the order makes a difference, but they consider a specific case of playing the bribery game and measuring social norms using the Krupka-Weber-method.

Treatment HIGHACTIVE

Welcome & Informed Consent

Thank you for participating in this study! The purpose of this study is to explore human decision-making.

This study is anonymous. We will not ask for your name or any information that will make you identifiable.

There is **no deception** in this study. Everything you see or read is true.

The study takes most participants less than 7 minutes to complete.

You will receive a fixed amount of £0.6 for your participation. You will have the chance to earn additional money at several points of the study depending on your behavior.

The risks to participating are no greater than those encountered in everyday life. Your participation in this study is completely voluntary, and you may refuse to participate or withdraw from the study without penalty. Compensation will be awarded upon completion of the entire study. If you have any questions, please contact us via Prolific. Please feel free to print or save a copy of this consent form.

Please tick the following box to be able to continue:

I have read and understood this consent form and wish to participate in this study.

The Die Roll Task

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). Only one number is the winning number and the decision is made in private and before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

Step Zero: Winning Number

In this step, please choose **one** winning number. It can be either 1, 2, 3, 4, 5, or 6.

This is your private decision. At a later stage, we will ask you whether your rolled number was equal to your winning number to determine your earnings.

Please remember your winning number.

Please click "next" when you are done with this step.

Step One: Die Roll

In this step, please roll a die and remember the result. **Important: Please use a regular six-sided die, please use only one die, please roll the die only once.**

You are free to use your own physical die. You can also go to a website like [this one \(click here\)](#) or any other die roll website you like.

Please click "next" when you are done with this first step.

Step Two: Reporting

Which number have you just seen as an outcome of your die roll?

- 1
- 2
- 3
- 4
- 5
- 6

Is your rolled number equal to your winning number?

- Yes
- No

Your additional payoff would be: ???

What do you think?

In a previous study, many participants have performed this exact same die roll task. What do you think: What was the percentage of people in this previous study who have reported their rolled number to be equal to their winning number?

Explanation: : Statistically speaking and if everyone was truthful, 17% should have reported their rolled number to be equal to their winning number (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: What was the percentage of people in a previous study who have reported their rolled number to be equal to their winning number?

Your guess (in %):

Do you need a reminder of the specifics of the die roll task? Here you can reread the text which you have already seen earlier.

[Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). Only one number is the winning number and the decision is made in private and before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

What do you think? Part II

In this previous study, some other participants have performed a **modified** version of the die roll task.

Please read through the description of the modified version. This is the description, which has been shown to the respective participants before the start of the game.

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment. If she reports a 5, she will receive **an additional £0.25**.

The statistical intuition to this question is the same as in the last question. Do you need a reminder of the statistical intuition? Here you can reread the information which you have already seen earlier. [Click here](#)

You will, again, earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: what was the percentage of people in a previous study who have played the modified game and who have reported a 5?

Your guess:

Do you need a reminder of the specifics of the **standard** die roll task in order to compare it to the **modified** version? Here you can reread the text which you have already seen earlier. [Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). Only one number is the winning number and the decision is made in private and before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

Results

You receive £0.6 for participating in this study.

You have reported to not have rolled your winning number, hence you receive an additional £0.0.

You have guessed that 5% of players in a previous study have reported their rolled number to be equal to their winning number in the standard game. The correct answer would have been 50%*.

Hence you receive an additional £0.00.

You have guessed that 33% of players in a previous study have reported a 5 in the modified game. The correct answer would have been 21%*.

Hence you receive an additional £0.40.

In total, you have earned £1.00.

We now kindly ask you to fill out a very short final questionnaire.

*Please note that these numbers are based on a relatively small sample size of 48 observations for the standard game and 48 observations for the modified game.

Questionnaire

Your earnings have been determined and the following question has no consequences for either your earnings, the approval of your submission or your rating on Prolific.co. During the study, we asked you about your beliefs: "What was the percentage of people in a previous study who have reported their rolled number to be equal to their winning number [who have reported a 5]?". Did you understand these questions?

- Yes
- No
- I can't remember
- None of the above

Questionnaire

How much do you agree with the following statement: 'I am an average person'?

- Strongly agree
- Agree
- Undecided
- Disagree
- Strongly disagree

Questionnaire

Your earnings have been determined and the following question has no consequences for either your earnings, the approval of your submission or your rating on Prolific.co. As promised, the outcome of your die-roll is not known to us. We, therefore, don't know if you have been truthful or not when reporting the outcome of your die-roll. For our research purposes, we would kindly ask you to voluntarily share this information by choosing one of the following options:

- I have lied to maximize my earnings
- I have told the truth and forgone some earnings
- The outcome of my die roll was the winning number, so I had no reason to lie
- None of the above
- I prefer not to answer

Questionnaire

The next question is about the following problem. In studies like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please enter turquoise as your answer to the next question. What is your favorite colour?

Colour

Questionnaire

We value your feedback!

Did you find anything unclear or misleading? Any technical issues? Any other feedback regarding any aspects of the study?

Would you like to explain your behavior in the study? Let us know!

Feedback

Treatment **LOWACTIVE**

Welcome & Informed Consent

Thank you for participating in this study! The purpose of this study is to explore human decision-making.
This study is anonymous. We will not ask for your name or any information that will make you identifiable.
There is **no deception** in this study. Everything you see or read is true.

The study takes most participants less than 7 minutes to complete.
You will receive a fixed amount of £0.6 for your participation. You will have the chance to earn additional money at several points of the study depending on your behavior.

The risks to participating are no greater than those encountered in everyday life. Your participation in this study is completely voluntary, and you may refuse to participate or withdraw from the study without penalty. Compensation will be awarded upon completion of the entire study. If you have any questions, please contact us via Prolific. Please feel free to print or save a copy of this consent form.

Please tick the following box to be able to continue:

I have read and understood this consent form and wish to participate in this study.

The Die Roll Task

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment.
If she reports a 5, she will receive **an additional £0.25**.

Step One: Die Roll

In this step, please roll a die and remember the result. **Important: Please use a regular six-sided die, please use only one die, please roll the die only once.**

You are free to use your own physical die. You can also go to a website like [this one \(click here\)](#) or any other die roll website you like.

Please click "next" when you are done with this first step.

Step Two: Reporting

Which number have you just seen as an outcome of your die roll?

- 1
- 2
- 3
- 4
- 5
- 6

What do you think?

In a previous study, many participants have performed this exact same die roll task. What do you think: what was the percentage of people in this previous study who have reported a 5?

Explanation: Statistically speaking and if everyone was truthful, 17% should have reported a 5 (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: what was the percentage of people in a previous study who have reported a 5?

Your guess (in %):

Do you need a reminder of the specifics of the die roll task? Here you can reread the text which you have already seen earlier.

[Click here](#)

Do you need a reminder of the specifics of the die roll task? Here you can reread the text which you have already seen earlier.

[Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment. If she reports a 5, she will receive **an additional £0.25**.

What do you think? Part II

In this previous study, some other participants have performed a **modified** version of the die roll task.

Please read through the description of the modified version. This is the description, which has been shown to the respective participants before the beginning of the task.

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). In this step, the participant does not report the decision to the researcher. Only one number is the winning number and the decision is made before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

The statistical intuition to this question is the same as in the last question. Do you need a reminder of the statistical intuition? Here you can reread the information which you have already seen earlier. [Click here](#)

Explanation: : Statistically speaking and if everyone was truthful, 17% should have reported their rolled number to be equal to their winning number (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will, again, earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: What was the percentage of people in a previous study who have played the modified game and who have reported their rolled number to be equal to their winning number?

Your guess:

Do you need a reminder of the specifics of the **standard** die roll task in order to compare it to the **modified** version? Here you can reread the text which you have already seen earlier. [Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment.

If she reports a 5, she will receive **an additional £0.25**.

Results

You receive £0.6 for participating in this study.

You have reported the number 5, hence you receive an additional £0.25.

You have guessed that 3% of players in a previous study have reported a 5 in the standard game. The correct answer would have been 21%*.

Hence you receive an additional £0.1.

You have guessed that 3% of players in a previous study have reported their rolled number to be equal to their winning number in the modified game. The correct answer would have been 50%*.

Hence you receive an additional £0.0.

In total, you have earned £0.95.

We now kindly ask you to fill out a very short final questionnaire.

*Please note that these numbers are based on a relatively small sample size of 48 observations for the standard game and 48 observations for the modified game.

Questionnaire

Your earnings have been determined and the following question has no consequences for either your earnings, the approval of your submission or your rating on Prolific.co. During the study, we asked you about your beliefs: "What was the percentage of people in a previous study who have reported their rolled number to be equal to their winning number?". And: "What was the percentage of people in a previous study who have reported a 5?". Did you understand these questions?

- Yes
- No
- I can't remember
- None of the above

Questionnaire

How much do you agree with the following statement: 'I am an average person'?

- Strongly agree
- Agree
- Undecided
- Disagree
- Strongly disagree

Questionnaire

Your earnings have been determined and the following question has no consequences for either your earnings, the approval of your submission or your rating on Prolific.co. As promised, the outcome of your die-roll is not known to us. We, therefore, don't know if you have been truthful or not when reporting the outcome of your die-roll. For our research purposes, we would kindly ask you to voluntarily share this information by choosing one of the following options:

- I have lied to maximize my earnings
- I have told the truth and forgone some earnings
- The outcome of my die roll was the winning number, so I had no reason to lie
- None of the above
- I prefer not to answer

Questionnaire

The next question is about the following problem. In studies like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please enter turquoise as your answer to the next question. What is your favorite colour?

Colour

Questionnaire

We value your feedback!

Did you find anything unclear or misleading? Any technical issues? Any other feedback regarding any aspects of the study? Would you like to explain your behavior in the study? Let us know!

Feedback

Treatment HIGHPASSIVE

Note: The consent page in the beginning and the questionnaire at the end are identical to the previous treatments and are therefore not shown again.

[Consent page]

The Die Roll Task

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). In this step, the participant does not report the decision to the researcher. Only one number is the winning number and the decision is made before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

What do you think?

In a previous study, many participants have performed this exact same die roll task. What do you think: What was the percentage of people in this previous study who have reported their rolled number to be equal to their winning number?

Explanation: : Statistically speaking and if everyone was truthful, 17% should have reported their rolled number to be equal to their winning number (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

We will provide you with the information about the behavior of one other participant: The first participant from the previous study has reported the number 4, the number which gave her a bonus.

What do you think: What was the percentage of people in a previous study who have reported their rolled number to be equal to their winning number?

Your guess (in %):

What do you think? Part II

In this previous study, some other participants have performed a **modified** version of the die roll task.

Please read through the description of the modified version. This is the description, which has been shown to the respective participants before the beginning of the task.

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment.

If she reports a 5, she will receive **an additional £0.25**.

The statistical intuition to this question is the same as in the last question. Do you need a reminder of the statistical intuition?

Here you can reread the information which you have already seen earlier. [Click here](#)

You will, again, earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: what was the percentage of people in a previous study who have played the modified game and who have reported a 5?

Your guess:

Explanation: Statistically speaking and if everyone was truthful, 17% should have reported a 5 (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

Do you need a reminder of the specifics of the **standard** die roll task in order to compare it to the **modified** version? Here you can reread the text which you have already seen earlier. [Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). In this step, the participant does not report the decision to the researcher. Only one number is the winning number and the decision is made before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

Results

You receive £0.6 for participating in this study.

You have guessed that 3% of players in a previous study have reported their rolled number to be equal to their winning number in the standard game. The correct answer would have been 50%*. Hence you receive an additional £0.0.

You have guessed that 4% of players in a previous study have reported a 5 in the modified game. The correct answer would have been 21%*. Hence you receive an additional £0.15.

In total, you have earned £0.75.

We now kindly ask you to fill out a very short final questionnaire.

*Please note that these numbers are based on a relatively small sample size of 48 observations for the standard game and 48 observations for the modified game.

[Questionnaire]

Treatment LOWPASSIVE

Note: The consent page in the beginning and the questionnaire at the end are identical to the previous treatments and are therefore not shown again.

[Consent page]

The Die Roll Task

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment.
If she reports a 5, she will receive **an additional £0.25**.

What do you think?

In a previous study, many participants have performed this exact same die roll task. What do you think: what was the percentage of people in this previous study who have reported a 5?

Explanation: Statistically speaking and if everyone was truthful, 17% should have reported a 5 (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: what was the percentage of people in a previous study who have reported a 5?

Your guess (in %):

Do you need a reminder of the specifics of the die roll task? Here you can reread the text which you have already seen earlier.

[Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment.
If she reports a 5, she will receive **an additional £0.25**.

What do you think? Part II

In this previous study, some other participants have performed a **modified** version of the die roll task.

Please read through the description of the modified version. This is the description, which has been shown to the respective participants before the beginning of the task.

For this study, it is important that you understand the structure of the "die roll task".

In step zero, every participant decides what her winning number will be (1, 2, 3, 4, 5 or 6). In this step, the participant does not report the decision to the researcher. Only one number is the winning number and the decision is made before the die is rolled.

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. She also reports if the rolled number is equal to her winning number. If the two numbers are equal, she will receive **an additional £2.5**. Otherwise, she will receive no additional earnings.

The researcher knows neither the rolled number nor the winning number. Hence, participants **do not have to fear any sanctions** from the side of the researcher or Prolific, regardless of what they report.

The statistical intuition to this question is the same as in the last question. Do you need a reminder of the statistical intuition? Here you can reread the information which you have already seen earlier. [Click here](#)

Explanation: : Statistically speaking and if everyone was truthful, 17% should have reported their rolled number to be equal to their winning number (16.67% to be more precise). However, some people might have lied in order to earn a higher pay-off. That means

... if you guessed 17% as an answer to this question, it would indicate that you believe that 0% of all participants have lied to their own benefit.

... if you guessed 100% as an answer to this question, it would indicate that you believe that 83% of all participants have lied to their own benefit.

You will, again, earn additional money for a good guess. You will receive £1 if you guess the correct number. This amount will be reduced by £0.05 for every percentage point deviation.

What do you think: What was the percentage of people in a previous study who have played the modified game and who have reported their rolled number to be equal to their winning number?

Your guess:

Do you need a reminder of the specifics of the **standard** die roll task in order to compare it to the **modified** version? Here you can reread the text which you have already seen earlier. [Click here](#)

For this study, it is important that you understand the structure of the "die roll task".

In the first step, every participant rolls a die **one time**. Only the participant herself knows the outcome of her die roll, i.e. the die roll is unrecorded by the computer and the **researcher does not receive this information**. Participants are free to roll their own physical die or use an online tool.

In the second step, every participant reports the outcome to the researcher. If the participant reports a 1, 2, 3, 4, 6, she does **not** receive any earnings in addition to the fixed payment. If she reports a 5, she will receive **an additional £0.25**.

Results

You receive £0.6 for participating in this study.

You have guessed that 21% of players in a previous study have reported a 5 in the standard game. The correct answer would have been 21%*.

Hence you receive an additional £1.0.

You have guessed that 33% of players in a previous study have reported their rolled number to be equal to their winning number in the modified game. The correct answer would have been 50%*.

Hence you receive an additional £0.15.

In total, you have earned £1.75.

We now kindly ask you to fill out a very short final questionnaire.

*Please note that these numbers are based on a relatively small sample size of 48 observations for the standard game and 48 observations for the modified game.

[Questionnaire]

Appendix C

IRB approval (ethical clearance)

The ethical standards of this work have been reviewed by the Dean's Office of the Faculty of Business, Economics and Social Sciences, University of Hamburg. Clearance was approved on September, 28, 2021. The respective document is available upon request.

Preregistration

A copy of the preregistration follows on the next pages. Note that the naming of the four treatments in the preregistration is slightly different compared to the main text.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Truth-telling and wishful thinking (#76432)

Created: 10/08/2021 04:25 AM (PT)

This is an anonymized copy (without author names) of the pre-registration. It was created by the author(s) to use during peer-review. A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

Lying leads to more pessimistic beliefs about the lying behavior of others.

3) Describe the key dependent variable(s) specifying how they will be measured.

Participants will answer the following question: "What do you think: What was the percentage of people in a previous study who have reported [a 5] OR [their rolled number to be equal to their winning number]?" Prior to asking the question, I explain the statistical intuition behind the question to the participants. This question results in the variable "beliefs about own condition". The question is incentivized.

4) How many and which conditions will participants be assigned to?

The experiment is built around the die rolling task (Fischbacher and Föllmi-Heusi 2013) which has been frequently used as a paradigm for truth-telling.

The study contains four experimental conditions: High Incentives Active (HIA), Low Incentives Active (LIA), High Incentives Passive (HIP), Low Incentives Passive (LIP).

I randomize participants into active and passive players. The former are exposed to a decision-making stage where they have the choice between lying and telling the truth, while the latter remain passive. This allows me to identify the causal effect of one's own past actions on beliefs. Further, subjects face one of two decision-making environments which provide different incentives to lie. These two conditions are designed in order to exogenously vary the share of truth-tellers among active players in the different subgroups.

I aim to have the same number of participants in each of the four conditions.

The study also contains one exploratory subtreatment (cf. section 8).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Outcome variable: "Beliefs about own condition".

To test the main hypothesis, I use a difference-in-difference approach (regression specification including an interaction term). The Diff-in-Diff estimator allows to analyze the effect of lying on beliefs while controlling for the varying incentive scheme, i.e. $(HIA - LIA) - (HIP - LIP)$. The passive conditions allow me to use the beliefs of "passive bystanders" in the same situation as benchmark. This specification includes standard controls for personal characteristics (such as age and gender).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Data points will be excluded from the analysis in four cases:

A participant did not complete the whole study.

A participant took an unrealistically small amount of time to finish the experiment: Less than 01:20.

A participant took an unrealistically large amount of time to finish the experiment: More than 5 standard deviations longer than the mean.

A participant has provided an unreasonable answer to the question regarding his beliefs about others (i.e. less than 1/6 have reported the winning number).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

I used the R package InteractionPowerR to conduct a power analysis for the relevant interaction term. Based on pilot data (03.08.21; 140 data points; data not included in the main analysis) and theoretical considerations, my study requires roughly 1200 observations to reach the conventional level of power of

80% (given $\alpha = 0.05$). All main effects and non-parametric tests (see below) should be adequately powered.

Some observations will be excluded (according to the exclusion criteria): Add 10% of observations: 1320 observations.

Roughly 180 observations will be collected for an additional exploratory treatment (cf. section 8).

Total amount of observations: 1500.

The data-collection via Prolific will be stopped automatically as soon as 1500 observations have been collected. Alternatively, I will end the data-collection when no new participants can be recruited for 72 hours.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

Additional analyses:

Non-parametric analysis of the main treatments

=> More pessimistic beliefs under HIA than under LIA ($LIA > HIA$): Wilcoxon rank-sum test

=> Slightly more pessimistic beliefs under HIP than under LIP ($LIP > HIP$): Wilcoxon rank-sum test.

Additional hypothesis: More people lie under high incentives than under low incentives (i.e. lying is context specific). The outcome variable is "share of liars". This is a prerequisite to be able to test the main hypotheses.

Analyzing which beliefs are more accurate: Comparing beliefs to actual behavior.

Additional dependent variable: "beliefs about other condition". Same as the main outcome variable but with respect to the other incentive condition.

Testing if more pessimistic beliefs are transferred to other scenarios.

Additional analysis based on the participants who have voluntarily admitted to have lied in during the experiment.

Exploratory subtreatment in treatment HIP: Providing information about the behavior of one other participant. This data will be used to compare the effect of (knowing) own behavior and the effect of knowing the behavior of one other participant.

Chapter 4

Advised by an Algorithm: Assessing Advice Quality with Informational Support and Reactions to Diverse Advice Quality

Authors: Jan Biermann, John J. Horton, and Johannes Walter

Abstract

Algorithmic recommendations are becoming increasingly common in many different situations. When humans take decisions with the help of algorithms, what information do they need to correctly assess the quality of algorithmic advice? Are humans able to effectively use the recommendations when the algorithm's advice quality fluctuates? We ask these two related questions in a setting in which an algorithm provides (partially biased) recommendations and the final authority to decide resides with the human. To answer these questions, we conduct an online experiment involving 1565 participants. In the baseline treatment, subjects repeatedly perform the same estimation task and receive an algorithmic recommendation without any information about the algorithm. We set out to test two feasible and apparent interventions intended to improve humans' judgment of the algorithmic advice, i.e., explanation of the algorithm and performance feedback. We find that while explaining the algorithm reduces adherence to the algorithmic advice, it does not improve human decision-making performance. Providing feedback in the form

of revealing the correct answer after each round, leads to a reduction in adherence to the algorithmic advice and an improvement in human decision-making performance. This study also investigates whether individuals can adjust their assessment of an algorithm when advice quality fluctuates. We find some evidence that individuals can assess algorithmic advice thoughtfully. Understanding the influence of external circumstances on algorithmic performance may be critical for humans to accurately assess the quality of advice.

Keywords: Human-algorithm decision-making, algorithmic advice, learning

JEL: C91, D79, D80, M21, O30

4.1 Introduction

Human decision-makers are increasingly being guided by algorithmic recommendation systems. This trend is evident in areas as diverse as healthcare, the judiciary, and e-commerce, where, respectively, physicians use algorithms to determine the most suitable treatments for patients, judges rely on them to support sentencing decisions, and pricing managers utilize them to strategically set discounts for products. Our paper focuses on the ability to assess algorithmic advice, as this skill is critical whenever a human holds the final decision-making authority. Our work provides experimental evidence in examining which types of informational support teach people to evaluate algorithmic advice. First, we study participants' response to an algorithm with a stable advice quality. Second, we analyze people's assessments under varying algorithmic performance.

To achieve the study's aims, we conduct an online experiment and ask 1565 participants to estimate the number of dots in an image. Our subjects receive algorithmic advice and are free to choose the extent to which – if at all – they want to incorporate this advice in their answer. The task is repeated for 16 rounds, using an image that has so many dots that counting them is infeasible. One important ongoing debate in the literature on human-AI decision-making focuses on humans' abilities to recognize and correct an algorithm if it malfunctions. We, therefore, study an algorithm that, under certain conditions, considerably underestimates the number of dots.

In the first part of the study, we test two interventions designed to improve a human's ability to optimally incorporate algorithmic advice. First, we provide participants with an explanation of how the algorithm arrives at its recommendation. Second, we reveal the solution to the prediction task after each round. This allows participants to better assess the algorithm's performance as well as their own ability. We also test a combination of both interventions. These interventions speak to two different concepts of how people could learn to assess the algorithm, namely learning by thought or learning by experience (Myagkov and Plott, 1997). In the former, participants could learn by receiving abstract

information about how the algorithm works, while in the latter, they could learn by directly experiencing the consequences of relying on the algorithm.

The recommendation quality of our algorithm is dependent on the interaction with the inputs (or environment) in which it operates. In the first part of our paper, we design the inputs (the dot images) in a way that leads to biased recommendations. In the second part, we also consider cases in which the same algorithm makes accurate predictions due to different kinds of inputs. Hence, our algorithm is not biased per se; rather, it delivers predictions of different qualities depending on the context. This mimics many real-world settings in which it is impossible for the developers of an algorithm to foresee all possible inputs or to estimate whether there would be a distributional shift in the input data. For example, Obermeyer et al. (2019) demonstrate that a health care algorithm advising physicians on which patient should receive a comparatively expensive treatment was biased for black patients, but not for white patients.

The focus of the second part of the paper is on evaluating individuals' ability to account for contextual influences on advice quality. Hence, using changing inputs, we expose subjects to a varying performance of the algorithm. Given that that algorithms are not equally suited to every environment, adjusting adherence to the algorithm based on the contextual circumstances becomes critical to achieving good outcomes.

Regarding our first intervention, we find that providing an explanation of the algorithm decreases participants' algorithm adherence, but it does not improve their guessing performance. We also find evidence that giving an explanation adversely affects the performance of some participants. Informing participants about what the true number of dots was at the end of the round also causes participants to follow the algorithm less, but in this case, it improves participants' guessing performance. Our analysis focuses on how the treatments influence outcomes compared to the baseline. We remain cautious about interpreting the absolute levels (of adherence to the algorithmic advice and the resulting performance) as these are strongly contingent on the task and context.¹

¹Existing work has, e.g., shown that individuals' willingness to trust algorithms depends on whether the task is (perceived as) objective (Castelo, Bos, and Lehmann, 2019) and whether the decision has a

Regarding the second part of our experiment, we find evidence that participants adjust their behavior to the changing environment and to the resulting varying quality of the algorithmic advice.² They put more weight on the advice if the algorithm produces good recommendations. Hence, they are, to some extent, capable of viewing the algorithm in a nuanced way, whereby they alternate between prioritizing reliance on the algorithmic advice and prioritizing their own assessment. In addition, adherence to algorithmic advice in an unfavorable environment does not depend on whether subjects encounter the algorithm only in that context or also in a favorable setting.

While AI has been shown to be capable of outperforming humans in various tasks (Lai et al., 2021; Kleinberg et al., 2018), there are also prominent cases illustrating that relying on dysfunctional algorithmic advice can have detrimental consequences (Angwin et al., 2016). The ability to recognize biased algorithmic advice becomes critical.³ Existing field experiments have provided evidence that people do not optimally incorporate AI advice in their decision-making and do not effectively update their beliefs about AI performance (Agarwal et al., 2023; Kim et al., 2024). This indicates that humans need further assistance when incorporating algorithmic advice in their decision-making.

In the field of computer science, several studies have empirically investigated different assistance interventions such as stating the accuracy of the algorithm (Yin et al., 2019) or providing confidence scores and local explanations (Zhang, Liao, and Bellamy, 2020; Alufaisan et al., 2021). It has also been shown that a slower response time of an algorithm can enhance the human evaluation of its performance (Park et al., 2019). From this angle, the study most closely related to our work is Green and Chen (2019), which reports experimental results on how people react to different aids, including two interventions

moral component (Bigman and Gray, 2018).

²In this part of the study, we provide our participants with the same two types of decision-making support as in the first part of the study, i.e., with explanation and feedback.

³Several existing studies have compared how decision-makers react to advice depending on whether the advice originates from an algorithm or from a human counterpart and they conclude that decision-makers' responses vary significantly (Önkal et al., 2009; Dietvorst, Simmons, and Massey, 2015; Prahla and Van Swol, 2017; Dietvorst, Simmons, and Massey, 2018; Logg, Minson, and Moore, 2019; Prahla and Van Swol, 2021; Sele and Chugunova, 2024). Examining the differences between human and algorithmic advice is therefore not the focus of our work. Rather, we acknowledge a fundamental difference as being established and ask the question: Given that the advice comes from an algorithm, which tools help decision-makers to better assess this advice?

that are similar to our treatments. However, Green and Chen (2019) do not investigate heterogeneous performance, nor do they discuss different ways of learning. Their study has a very different experimental set-up and task, i.e., pretrial release and financial lending, than ours.

Related to our investigation of how people react to varying algorithmic performance, several studies have looked at how people behave when they notice that the algorithm makes mistakes. Dietvorst, Simmons, and Massey (2015) have shown that people tend to abandon algorithmic advice after recognizing algorithms' errors. Several subsequent studies have explored different nuances of this seminal finding (e.g., Prahla and Van Swol, 2017; Dietvorst, Simmons, and Massey, 2018; Jung and Seiter, 2021; Zhang and Gosline, 2022; Reich, Kaju, and Maglio, 2023).

A number of studies in the field of decision-making under risk are related insofar as they explore the two types of learning that our work examines through the interventions. Myagkov and Plott (1997) are the first to distinguish between learning by thought alone (i.e., without feedback) and learning by thought and experience (i.e., with feedback). Several studies have empirically analyzed the difference between these learning types, providing mixed evidence as to whether learning by thought alone is effective.⁴ Building on these findings, we apply this framework to our domain of interest, i.e., to the ability of humans to evaluate algorithmic advice. While the importance of feedback and directly experiencing the consequences of using algorithms in decision-making has been discussed in other domains, providing feedback has not received much attention from policymakers in the context of AI regulation. For example, while article 14 in the EU's AI Act (European Commission, 2021) discusses several measures to ensure human oversight (e.g., appropriate training and explanation), it neither implicitly nor explicitly mentions feedback.

We contribute to the existing literature in two major ways. First, we test interventions designed to improve people's assessment of algorithmic advice. Importantly, our inter-

⁴van de Kuilen and Wakker (2006) and van de Kuilen (2009) find that learning by thought and experience leads to more rational behavior, while learning by thought alone does not. Conversely, Hey (2001), Birnbaum and Schmidt (2015), and Nicholls, Romm, and Zimper (2015) report that learning by thought alone is sufficient to increase rational behavior over time.

ventions are informed by the existing work on decision-making under risk. We introduce the concept of learning through thought vs. learning through experience to the literature on empirical human-AI decision-making. In doing so, we ensure that the debate on how people learn about algorithms becomes more nuanced.

Second, we investigate the extent to which individuals can skillfully adjust their evaluations of the algorithm in response to varying levels of advice quality. We create a scenario in which algorithmic performance fluctuates based on external factors (i.e., we analyze a stable algorithm in a changing setting resulting in a varying performance). Our contribution is to illustrate that subjects do not inevitably discard the algorithm after seeing it make poor predictions. If the low-quality advice is caused by the algorithm being unsuitable for a certain setting, humans will continue to use it under different circumstances.

Further, our study is innovative with regard to the experimental task we employ. To the best of our knowledge, we are the first to investigate algorithm-supported decision-making through the dot-guessing task. This task is attractive for several reasons. It is accessible to lay people without specialized expertise. The general skill of counting dots is one to which all humans have access (except under certain medical conditions). Further, by employing this task we are able to vary the quality of algorithmic advice by altering external conditions (i.e., the distribution of dots) while keeping the functionality of the algorithm unchanged.

The remainder of this paper proceeds as follows: We subdivide the main part of the paper into two parts. Section 4.2 investigates the effects of providing informational resources (part 1). Section 4.3 examines our treatments employing varying performance (part 2). Both parts contain subsections explaining the experimental design and presenting the results. Section 4.4 discusses the findings of both parts. Section 4.5 gives the conclusion.

4.2 Part I: Informational Support

4.2.1 Experimental Design

Experimental Task and Algorithm

The central component of our experiment is the dot-guessing task. Subjects see images showing a large number of blue dots and are asked to guess how many dots they think each of the images contains. Examples of dot images are shown in figure 4.2. The number of dots in the images is chosen randomly and varies between 942 and 3084 dots. Participants have 60 seconds to make their guesses, making it infeasible to count the dots in the image.⁵

Every round of the experiment consists of three stages: In the first stage, participants see the image for the first time and submit their guesses. In the second stage, subjects see the same image again and additionally receive an algorithmic prediction of the number of dots, before submitting a new guess. We call the two entries the subjects make initial guess $guess_i$ and revised guess $guess_r$, both of which are incentivized. In the third stage, participants can see their guesses $guess_i$ and $guess_r$ and some additional information depending on the treatment. They do not take any action at this stage. Every subject plays 16 rounds, each round including a new image and a new recommendation. In each round, all participants see the same image (i.e., the same number of dots) and receive the same recommendation. We employ a between-subject design and randomize our participants on an individual level.

The algorithm we employ samples three subareas from the given image, calculates the average of dots within these subareas, and extrapolates this average to the entire surface of the image. We introduce a bias to the predictions of the algorithm in the following way: We restrict the sampling of the algorithm to the outer edges of the image and in addition we choose a triangular distribution with a denser center and fewer dots at the

⁵Our task is rooted in the tradition of Galton (1907). His research has produced the “wisdom of the crowd” finding and involved a contest in which people guessed the weight of a butchered ox.

edges. As a result, the algorithm systematically underestimates the number of dots in the entire image.⁶ See figure 4.2 for further illustrations.

Description of Treatments

The treatments vary along two dimensions: explaining how the algorithm arrives at its prediction and revealing the true answer after the revised guess has been recorded. The design of our interventions speaks to whether humans learn by thought only or by thought and experience (Myagkov and Plott, 1997) and is in this respect related to several empirical studies which apply this concept to test expected utility-theory (Hey, 2001; van de Kuilen and Wakker, 2006; van de Kuilen, 2009; Birnbaum and Schmidt, 2015; Nicholls, Romm, and Zimmer, 2015).

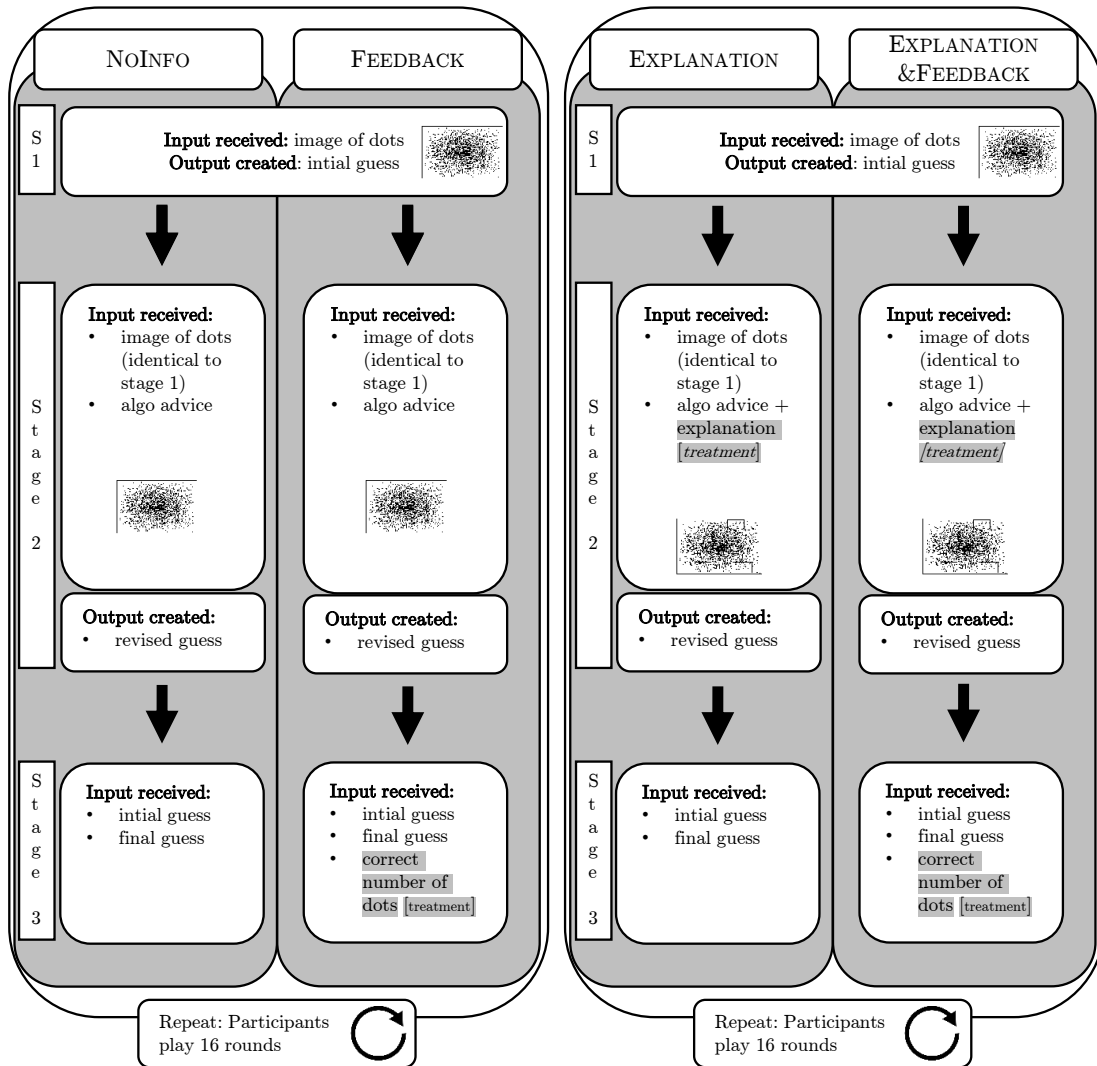
In the treatments EXPLANATION and EXPLANATION&FEEDBACK, participants receive an explanation of how the algorithm works. This includes a visual and a textual component. Participants see the image of dots overlaid with squares indicating the subareas the algorithm samples from (cf. figure 4.2). We also inform them in writing that the algorithm counts the number of dots in the visualized squares and then predicts the total number of dots in the entire area based on this sampling. Hence, we do not explicitly state to participants that the predictions are biased, but we deem that our explanation provides the necessary information to comprehend that the algorithm vastly underestimates the number of dots.⁷ See appendix for experimental interfaces containing the exact wording.

Furthermore, in the treatments involving feedback, participants receive information about the correct number of dots in the image at the end of every round. In treatments without feedback, participants never find out the correct answer to the task. Seeing the solution provides an opportunity to assess the performance of the algorithm in a specific round. It also opens the chance to learn about one's own performance.

⁶Images with uniform distributions will be employed in the second part of the experiment resulting in unbiased algorithmic predictions.

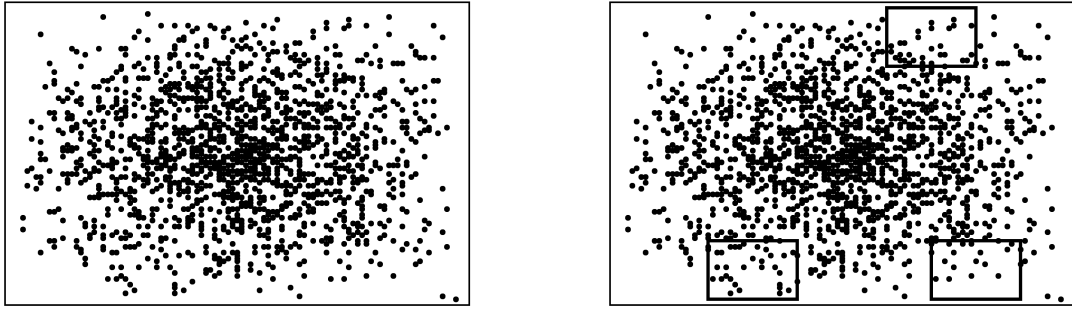
⁷This information is provided in the second stage of each round, i.e., when participants state their revised guess.

Figure 4.1 Visualization experimental design part I



To sum up, the two dimensions result in the four treatments NOINFO, EXPLANATION, FEEDBACK, EXPLANATION&FEEDBACK. These treatments enable us to analyze which type of information empowers humans to effectively assess the quality of algorithmic advice and how they influence task performance. The experimental design is visualized in figure 4.1.

Figure 4.2 Functioning dot guessing algorithm



(a) Example dot image in treatments without explanation
(b) Example dot image in treatments with explanation

Notes: The algorithm arrives at its prediction by first randomly sampling three subareas from the edges of each image and counting the number of dots within each subarea. It then calculates the average number of dots over areas and projects this average to the entire image. Importantly, the algorithm always samples the three subareas at the edges of a given image, never from the center. Through the combination of triangular dot distribution and sampling being restricted to the edges, we introduce a bias in the algorithm prediction. The image in panel (a) is an example of a dot image that all participants see in the first stage of the experiment. Panel (b) shows the same image but this time overlaid with the rectangular subareas from which the algorithm samples dots. Only participants in the treatments with explanation see this image from panel (b) in the second stage, complementing the verbal algorithm explanation.

Payment Scheme and Experimental Procedure

Our subjects receive a flat fee of \$0.9 for completing the study. Guesses are incentivized as follows: Participants receive \$0.15 for a perfect guess, and this bonus is diminished by \$0.0002 for each point difference, resulting in a bonus if a participant's guess is within the range of ± 749 dots from the true answer.⁸ The experiment contains 16 rounds and each of the rounds involves two incentivized guesses (initial and revised). Therefore, subjects could earn a maximum of \$4.8 bonus payment in addition to the flat fee.

We conducted our online experiment in December 2021. The experiment was developed using the software oTree (Chen, Schonger, and Wickens, 2016). We recruited our subjects via Amazon's crowd-working platform Mechanical Turk (MTurk). All of them are based

⁸The first visual impact of the images might be discouraging for certain participants. Hence, we deliberately selected a broad range (almost 1500 dots) within which participants receive a bonus. This is done to encourage our subjects to maintain an aspiration for receiving a bonus and to motivate them to strive for accurate estimations.

in the US, have completed at least 500 tasks on MTurk, and have an approval rate of at least 95%. We conducted five sessions (two sessions with 200 and 3 sessions and 400 participants). Our data contain 1565 observations in total. 1263 of them are relevant for the four treatments in part I and an additional treatment containing 302 observations will be introduced in part II. On average, participants took 14 minutes and 18 seconds to complete the study and earned \$2.33. This translates to a hypothetical hourly wage of \$9.82.

4.2.2 Data

The main data we elicit from our subjects are their guesses with respect to the number of dots in the image they see. When eliciting these guesses, we do not set an upper bound (e.g., by employing a slider) as such an upper bound would serve as an orientation point for some of our subjects. As a result, participants can enter very high numbers, and in fact, some choose to do so. We, therefore, see large outliers in our distributions. Three approaches are common to address the issue of outliers in the data: top-coding, winsorizing, and taking the natural logarithm. We employ the latter method. Similarly, although some participants state a very low guess for the number of dots, including 0, we do not exclude these guesses at the lower end of the distribution either. This approach has the advantage of not requiring us to exclude any observations from our analysis. Instead, we can show that excluding very low guesses does not change our main results. For a more in-depth discussion of this issue cf. the appendix.

Further, since the number of dots and the algorithmic advice changes every round, examining (the logarithm of) the guesses at their face value would have little meaning. We are rather interested in the relation between guesses and i.) the algorithmic advice or ii.) the correct number of dots. Therefore, we focus on two main outcome variables throughout the paper: algorithm adherence, calculated as $|\log(algo) - \log(guess)|$, and guessing performance, calculated as $|\log(truth) - \log(guess)|$. Tables 4.1 - 4.6 in the appendix give an overview over these main outcome variables.

Various parts of the analysis are based on a comparison among treatments in which case we pool all rounds together. We recognize that the guesses of each individual are not independent of each other. We, therefore, preprocess the data by calculating the mean of the values of interest (e.g., distance to algorithmic recommendation) for all 16 rounds for each individual.

4.2.3 Results

In this section, we report results regarding how different aids to better assess the algorithm influence performance and algorithm adherence. More specifically, we are interested in whether revealing the correct answer at the end of the round or providing an explanation of the mechanics of the algorithm can help subjects learn from the interaction with the algorithm. First, we analyze the effects on algorithm adherence by examining the distance between the average revised guess and the algorithmic prediction for each treatment. Figure 4.3a illustrates that feedback reduces algorithm adherence compared to the baseline treatment.⁹ The same is true for explanations with an even stronger effect. We observe the strongest reduction of algorithmic adherence in the treatment EXPLANATION&FEEDBACK.

Result 1a: *Explanation reduces algorithm adherence (medium effect).*

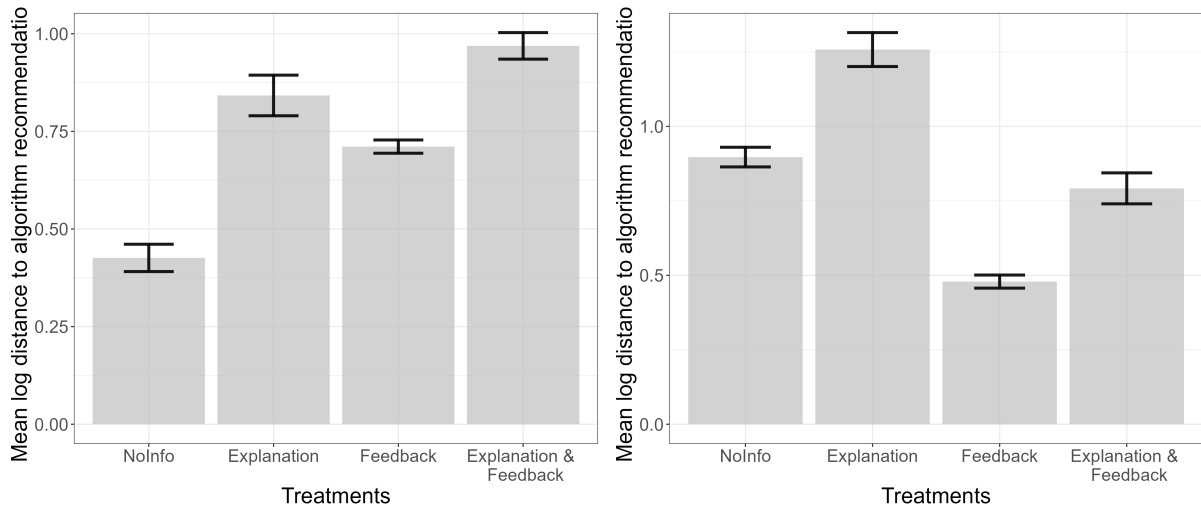
Result 1b: *Revealing the truth reduces algorithm adherence (weakest effect).*

Result 1c: *Combining explanation and the revealing truth reduces algorithm adherence (strongest effect).*

Knowing that both interventions reduce algorithm adherence, we now turn to the question of how the treatments influence guessing performance. We hence examine the distance between the revised guess and the correct answer. We present the results in figure 4.3 and table 4.3 in the appendix. Based on figure 4.3, providing the explanation increases the

⁹Table 4.2 in the appendix contains more detailed, numeric information regarding algorithm adherence.

Figure 4.3 Mean distance to the algorithm and the true number of dots by treatment



(a) Mean distance to the algorithm recommendation by treatment. **(b)** Mean distance to the true number of dots by treatment.

Notes: The bar graph in panel (a) illustrates the treatment effects on algorithm adherence (mean distance of the revised guesses to the algorithm recommendation per treatment). The numerical treatment effects on algorithm adherence can be found in table 4.2 in the appendix. The bar graph in panel (b) illustrates the treatment effects on guessing performance (mean distance of the revised guesses to the true number of dots per treatment). The numerical treatment effects on guessing performance can be found in table 4.3 in the appendix. The bar plots also include the standard errors around the mean. For the outcome distance to the algorithm recommendation, all three treatments are significantly different from the baseline NOINFO treatment on a 1% level based on an independent two-sample t-test. For the outcome distance to the true number of dots, only treatment EXPLANATION&FEEDBACK is *not* significantly different from the baseline on a 1% level. We pre-process the data by taking log values and calculating the mean over all 16 rounds for each individual. For more details on data pre-processing see section 4.2.2.

average distance to the true number of dots compared to NOINFO, i.e., the explanation makes participants perform worse.¹⁰ While this result is striking, we remain cautious about over-interpreting this finding. First, we provide a robustness check in the appendix and find that the negative effect does not disappear after excluding extremely small guesses, but is starkly reduced. Second, our experiment is not designed to uncover the underlying mechanisms of this result.¹¹ For these two reasons, we err on the side of caution in the main body of the paper in proposing the interpretation that “explanation does not improve performance.”

¹⁰Compare section 4.2.1 for more information on the content of the explanation.

¹¹We, therefore, can not offer any causal experimental insights, but we provide a discussion of potential driving forces of this result in the appendix.

In contrast, revealing the correct answers improves guessing performance. Remarkably, the effects of these two treatments have approximately the same size (and opposite directions). As a result, in treatment EXPLANATION&FEEDBACK, the two effects appear to cancel each other out. Hence, the average performance under EXPLANATION&FEEDBACK is statistically indistinguishable from NOINFO (t-test is not significant on a 5%-level).

Result 2a: *Explanation does not improve performance (and possibly hurts performance).*

Result 2b: *Feedback improves performance.*

Result 2c: *Combining explanation and feedback does not significantly change performance compared to the baseline treatment.*

In the appendix, we also provide an additional analysis of the baseline treatment. It suggests that without any additional aids, most subjects are not capable of assessing the quality of the algorithm. One might therefore consider algorithmic advice to be a credence good. However, we do not emphasize this analysis because such results are highly dependent on the specific task and setting. Additional discussions and interpretations of the results from this section follow in section 4.4.

4.3 Part II: Diverse Advice Quality

4.3.1 Experimental Design

The second part explores how people react to varying performance of the algorithm. To this end, we introduce an additional treatment which we call VARYINGQUALITY. The experimental setup is largely the same as in the first part: Participants play the dot guessing game for 16 rounds, they submit initial guesses, observe an algorithmic recommendation, and then submit their revised guess. The appearance of the interfaces and the incentive structure are identical to the first part. Importantly, in this treatment our

participants receive both informational resources: the explanation of the algorithm and the solution of how many dots were in the image at the end of each round as in EXPLANATION&FEEDBACK from the first part. Therefore, we use EXPLANATION&FEEDBACK as a benchmark in this section.

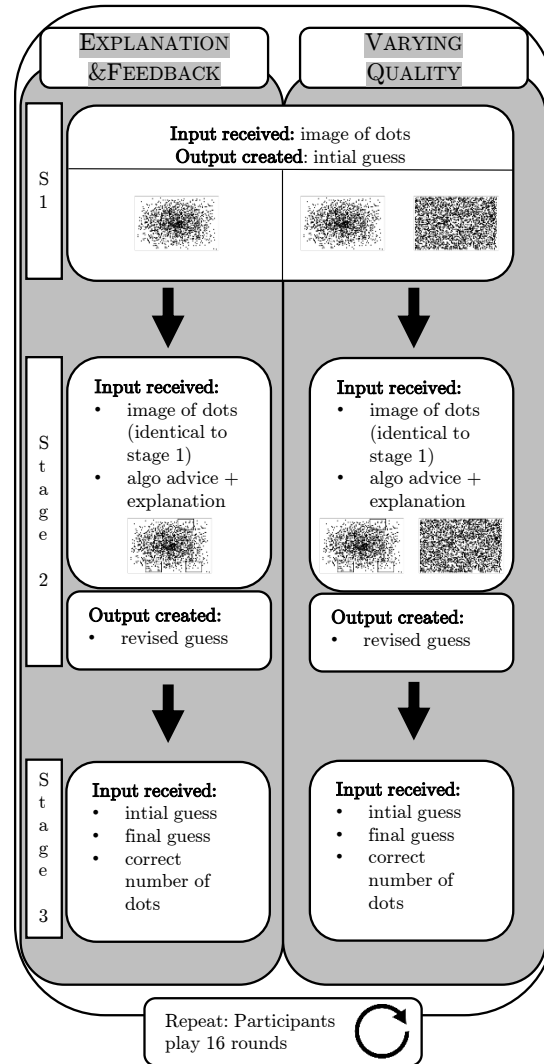
The difference to EXPLANATION&FEEDBACK is that participants in VARYINGQUALITY see only 8 images in which the dots follow a triangular distribution. The other 8 images show dots following a uniform distribution. Each participant in this treatment sees an image where the dots are triangularly distributed in every odd round and an image with uniformly distributed dots in every even round. The 8 dot images with the triangular distribution are the same as the first 8 images in the four treatments in EXPLANATION&FEEDBACK. Figure 4.5 shows two exemplary images for each type of distribution.

Changing the distribution has consequences on the performance of the advising algorithm. Since the algorithm always samples from the edges of an image, it draws a biased sample in case of a triangular distribution. In contrast, when the dots are uniformly distributed, the sampled areas are representative of the whole image and this results in precise algorithmic predictions. This setting mimics many real-world scenarios the in which algorithm's functionality remains consistent; however, the varying settings result in the algorithm's techniques yielding either accurate or inaccurate predictions. The experimental design is visualized in figure 4.4.

4.3.2 Results

Do our participants disengage from the algorithm after seeing it perform poorly? Or do they recognize when to rely on the algorithm and when to ignore it? Figure 4.6 shows the distance of the revised guesses to the algorithm recommendation averaged over three different sets of rounds: The left bar shows shows the average over all 16 rounds in EXPLANATION&FEEDBACK which is our benchmark (algorithmic advice is of poor quality). The bar in the center shows the average over the 8 rounds in which the algorithm's advice is of high quality in VARYINGQUALITY. The right bar shows this

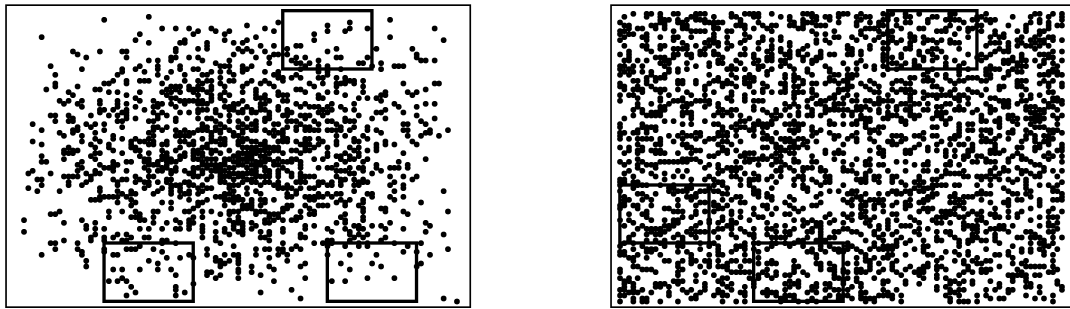
Figure 4.4 Visualization experimental design part II



distance averaged over the other 8 rounds in which the algorithm's advice is of poor quality in VARYINGQUALITY.

The results illustrate that in rounds with high advice quality (uniform distribution), subjects follow the algorithm more closely than in rounds with low advice quality (triangular distribution). These differences are significant on a 1%-level (independent two-sample t-tests). Hence, our participants appear to continue to engage with the algorithmic advice (of varying quality) even after encountering some low-quality advice. This may seem at odds with previous literature showing that people do not forgive algorithmic errors (Dietvorst, Simmons, and Massey, 2015). We interpret this apparent discrepancy in the

Figure 4.5 Uniform and triangular distribution of dots

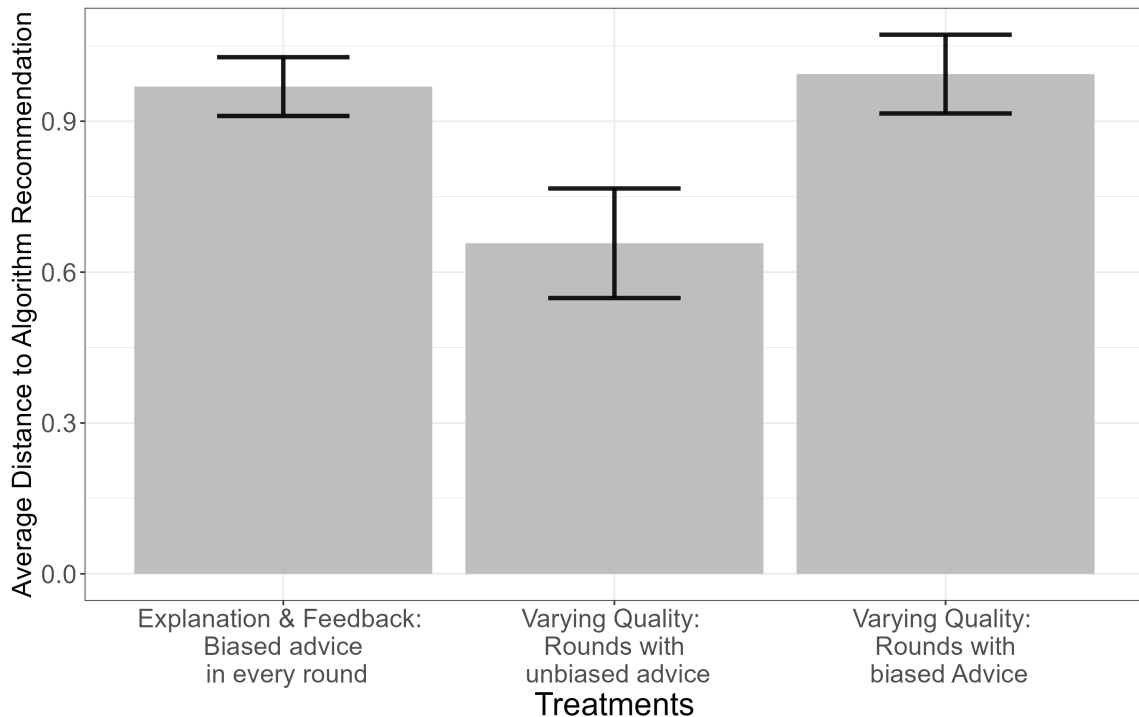


(a) Triangular dot distribution with boxes

(b) Uniform dot distribution with boxes

Notes: Participants in VARYINGQUALITY see images alternating between triangularly distributed dots as in panel (a) and uniformly distributed dots as in panel (b).

Figure 4.6 Mean distance to the algorithm for EXPLANATION&FEEDBACK and VARYINGQUALITY



Notes: All three bars show the average distance of the log revised guesses to the algorithm recommendation, including the standard errors around the mean. The leftmost bar does so for EXPLANATION&FEEDBACK. The middle bar shows this result for VARYINGQUALITY, but only for the 8 rounds in which participants received *unbiased* recommendations. This treatment is significantly different from each of the other two treatments on a 1%-level based on independent two-sample t-tests. The rightmost bar shows this result for VARYINGQUALITY, but only for the 8 rounds in which participants receive *biased* recommendations. For more details on data pre-processing see section 4.2.2.

discussion section.

Result 3: *People follow algorithmic advice more in environments where the algorithm gives high-quality advice than in environments where it gives low-quality advice.*

Further, there is no statistically significant difference between distance to the algorithm when advice quality is poor in VARYINGQUALITY compared to EXPLANATION&FEEDBACK. In other words, there are no spill-overs of trust from the high-quality advice rounds to the low-quality advice rounds in VARYINGQUALITY. Participants' assessment of the algorithm in poor-advice rounds is not influenced by experiencing the algorithm in a favorable setting.

Result 4: *Seeing the algorithm perform well in a favorable setting does not increase trust in the algorithm in other settings (i.e., there are no positive spill-overs).*

4.4 Discussion

4.4.1 Informational Support

Following the concept of Myagkov and Plott (1997), we examine whether experience is a necessary component for decision-makers to learn and evaluate the quality of algorithmic support systems. Existing empirical literature from different contexts has provided inconclusive evidence on whether or not learning by thought alone is effective.¹² Addressing this question in the context of algorithms is of vital importance. Consider high economic impact decisions that managers have to make. Learning through thought is always

¹²Compare Hey (2001); van de Kuilen and Wakker (2006); van de Kuilen (2009); Birnbaum and Schmidt (2015); Nicholls, Romm, and Zimper (2015).

available, as the firm can simulate a series of hypothetical decisions. Learning through experience is not available in some settings, e.g., when a decision occurs only once, the consequences are unclear or only occur after a certain time lag. However, the results from our experiment indicate that learning by thought is not sufficient and experiencing the consequences of (dis-)trusting the algorithm is required in order to improve assessment. Hence, the practical implication of this is that providing timely feedback can help improve decisions and it is worthwhile to strive to implement such feedback in situations where it is feasible.

Providing an explanation increases the participants' average distance to the algorithm compared to the baseline. This suggests that the explanations help participants recognize that the algorithmic predictions are biased (in the case of triangular distributions). At the same time, explanation does not improve performance. This is arguably because – after recognizing that algorithmic predictions are biased – participants still don't have a better prediction at their disposal. Taken together, the effects on performance and algorithm adherence can be seen as evidence of the ability to detect a bias and the inability to correct a bias.¹³

The absence of a positive impact of explanation on performance is particularly striking. We, therefore, present an alternative specification as a robustness analysis in the appendix by excluding unreasonably small guesses. While the negative effect on performance is smaller in this specification, it does not disappear completely. Nevertheless, given the reduction in effect size, we err on the side of caution and conclude that explanation does not improve performance (and possibly even has an adverse effect).

We discuss potential interpretations for why explanation does not improve performance and why it leads to some unreasonably low guesses in the appendix. Overall, it appears that the explanation we have provided has displeased or confused some participants and did not have a positive effect on the average performances. Such heterogeneous treatment effects can also be expected in many real-life applications: While some decision-makers

¹³In principle, explanation could have led to better performance because participants realized that the algorithmic advice is biased, but can be used as a lower bound for a better estimation. However, we don't see evidence for this.

will benefit from such explanations, others will ignore them in the face of their busy workday, and still others might feel frustrated when having to read them. This highlights the importance of providing such descriptions in a way that is appropriate for the target user. Despite discussing the potential drivers of this finding, we do not draw any definitive conclusions about the underlying mechanisms since our experiment was not designed for such analysis. Further investigating the patterns of this adverse effect is an interesting direction for future research.

Under FEEDBACK participants move further away from the algorithm, but in this treatment, they also move closer to the true number of dots. One potential mechanism of how subjects arrive at their guess is the following: They start with an *orientation point* and adjust this amount of dots based on their judgment. Examining figures 4.15 to 4.18 (in the appendix) provides some suggestive evidence that this is indeed an important mechanism. The distribution of initial guesses in the first round is flat. This illustrates that initially subjects are uncertain with regard to their own guess. In addition, participants strongly react to the algorithmic advice. The specific number of dots recommended by the algorithm gives the participants a potential orientation point to use.

The algorithmic advice is an orientation point that is always available when revising the guess. Importantly, in the treatments where people receive feedback, an alternative orientation point is provided: the true answer from the previous round(s). Therefore, providing subjects with the opportunity to better access the quality of the algorithmic advice is not the only channel through which the feedback in our experiment could influence participants.¹⁴

The concept of the orientation point might be considered particularly important in the light of the fact that figures 4.15 to 4.18 suggest that participants are remarkably often close to the true answer in treatments where they receive feedback. It appears that participants are strong in taking the true number of dots as an orientation point, estimating the relative change of dots in the next round, and delivering a good estimate for the current round.

¹⁴Further, it may also impact participants in a third way: providing feedback about their own abilities.

Treatment EXPLANATION&FEEDBACK shows an interesting cumulative result: Since providing an explanation and revealing the truth both increase the distance to the algorithm, their combination does as well. And as these treatments individually have opposed effects on guessing performance, their combination seems to annihilate any effect. Guessing performance remains unchanged compared to the baseline. This illustrates that practitioners must be careful when considering tools to improve decision-making. Generally, one cannot simply assume that the more help available to the decision-maker, the better.

Despite the algorithmic predictions being biased (given a triangular distribution), it is an empirical question whether it delivers better predictions than our participants. Answering it is relevant because it determines whether participants would have enhanced their performance by incorporating algorithmic advice into their revised guesses. Generally speaking, there is a sizable proportion of people who have performed better as well as worse than the algorithm. Consequently, there is notable heterogeneity in whether increased or decreased adherence to the algorithm would have resulted in better revised guesses. On average, participants would have gained from assigning a positive weight to the algorithmic advice in most rounds in NOINFO (algorithmic prediction superior in 14/16 rounds) and EXPLANATION (algorithmic prediction superior in 13/16 rounds). Conversely, this is not observed in FEEDBACK (participant guess superior in 14/16 rounds) and EXPLANATION&FEEDBACK (participant guess superior in 12/16 rounds). We provide the corresponding analysis in the appendix.

Another important aspect is whether our interventions show their effect immediately or if some repeated interaction is required for the effect to unfold. Overall, there appears to be no pattern that evolves. The interventions do not seem to require warm-up time. One exception is the first rounds of EXPLANATION. Participants move closer to the algorithm in the first two rounds after receiving the advice. In round three, the average distance remains the same. Starting from round four, participants always move further away from the recommendation (cf. figures 4.15 to 4.18). This suggests that participants must see the explanation multiple times before the effect starts unfolding.

4.4.2 Diverse Advice Quality

Our results on varying algorithmic performance show that participants adhere more to algorithmic advice in rounds with good performance and less in instances of poor performance. This may appear in contrast to Dietvorst, Simmons, and Massey (2015), who present evidence indicating that subjects tend to abandon the algorithm upon observing errors.

Most likely, the difference between the two setups is due to the fact that our participants have the opportunity to comprehend the algorithm’s suitability for specific settings. We provide our participants with informational resources, aiding their understanding that the algorithm isn’t inherently flawed but varies in performance based on the setting. In Dietvorst, Simmons, and Massey (2015), participants do not have the opportunity to understand the causes of the observed mistakes.¹⁵

In this sense, one of our contributions is to investigate the circumstances under which the pivotal result of non-forgiveness towards algorithms holds and to provide a context for it. Another way to read the Dietvorst, Simmons, and Massey (2015) result is that they document a negative spillover with regard to trust in algorithms: After encountering low-quality advice, individuals tend to distrust the algorithm in other cases. Our study addresses whether there are spillovers in the opposite direction. We find that after witnessing the algorithm perform well in a favorable setting, people do not trust it more in other cases.

Bringing together our results from both parts of the paper, one can say that FEEDBACK improves overall performance, while EXPLANATION does not. Hence, learning by thought alone is insufficient to improve performance in our context. Given both informational resources, subjects are capable of distinguishing when to follow the algorithmic advice and when to leave it aside. The results of the two parts of our paper can be seen as evidence that the question of whether people can judge the quality of algorithmic advice

¹⁵The same holds true for relating our study with Bao et al. (2022) who have shown that people are reluctant to follow advice from algorithms that appears inconsistent.

is highly nuanced and dependent on many factors. On the one hand, our results suggest that, provided with our explanation, subjects have a hard time assessing the capacities of the algorithm in relation to *their own capabilities*. On the other hand, given explanation and feedback, subjects seem to be able to adequately assess algorithmic advice in relation to *advice from previous rounds* when performance varies.

4.4.3 Caveats and Potential Extensions

Some important caveats are in order. First, we exercise caution in generalizing our findings to all scenarios involving human-machine interaction. We recognize the contextual specificity of certain behaviors and focus on interpreting differences between treatments rather than the absolute levels of our variables of interest. Our hope is that future research will explore how our findings translate to different settings.

Second, our emphasis is on analyzing how algorithmic advice influences behavior outcomes (rather than self-reported measures). Our results hint at the fact that subjects hold (overly-)optimistic beliefs regarding the algorithm’s capabilities when we do not provide any information regarding the algorithm. Further, we argue that reliance on algorithmic advice is driven by the interplay of two types of beliefs: beliefs about algorithmic performance and beliefs about one’s own capabilities. A potentially fruitful avenue for future research involves eliciting and correlating beliefs with behavioral outcomes to better comprehend the underlying mechanisms.

Third, it would be interesting to conduct an additional treatment that includes feedback after each round but does not provide algorithmic advice. This would make it possible to separate the different components of providing feedback. In the current setup, by being informed about the correct answer, subjects can infer information about the algorithmic performance as well as their own performance (and also receive an orientation point).

Fourth, the current study explores at the assessment of varying algorithmic performance while providing both feedback and an explanation. Exploring how reactions differ when

individuals receive only explanations or only feedback in this setting could provide further insights.

Fifth, our study is not designed to pinpoint the reasons for why providing an explanation does not improve performance. While we have suggested some mechanisms, a more in-depth analysis of this phenomenon would be a valuable extension of our work.

4.5 Conclusion

We design an experiment in which subjects are asked to guess the number of dots they see in an image while receiving advice from an algorithm. We test a set of interventions - providing an explanation of the algorithm, revealing the solution *ex post*, or both - and ask whether they can improve the ability to assess the algorithm.

Our results suggest that providing an explanation of the algorithm does not improve performance and may possibly even hurt performance. Revealing the truth *ex post* on the other hand does benefit people's performance. We are cautious about attributing this solely to participants learning about the algorithm's quality. In this treatment, subjects also have the chance to learn about their own performance and they receive a benchmark for calibrating future guesses. These results suggest that learning by thought appears to be insufficient to improve the assessment of algorithmic quality. In our setting, people need to experience the consequences of their decisions and hence feedback is required.

Our study further considers a setting of varying algorithmic performance due to changing circumstances. We argue that people do not abandon algorithms when errors occur if they can comprehend the reasons behind the errors. Subjects appear to have some ability to understand the strengths and weaknesses of the algorithm and to use it in cases where it proves beneficial.

These findings suggest several practical recommendations. First, in organizational settings, if concrete feedback about previous decisions can be provided and the decision

environment does not fundamentally change, managers should seek to disclose the outcome of past decisions. Second, our study indicates that explanations concerning how an algorithm functions must be provided with caution, as they do not necessarily improve human assessment of algorithm quality. This finding is in line with the results from the literature on algorithm explainability: Its ability to improve decisions also depends on various factors, and explanations are not a panacea. Third, there is hope that decision-makers can be trusted to assess imperfect algorithms, provided they receive sufficient informational resources for learning and have the opportunity to comprehend the reasons behind algorithmic errors.

Acknowledgments

Valuable discussions with the following people have greatly improved this paper: Robert Dur, Kris Johnson Ferreira, Ben Green, Adrian Hillenbrand, Rudi Kerschbamer, Lydia Mechtenberg, Dominik Rehse, and Marco Schwarz. We also want to thank the participants of TUHH Institute for Digital Economics Seminar 2021, ZEW Digital Economy Seminar 2022, YEM 2022, ASFEE 2022, UHH Collective Decision-Making PhD Seminar 2023 and Innsbruck Spring Summit on (Un)Ethical Behavior in Markets 2023.

Bibliography

- Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. “Combining human expertise with artificial intelligence: Experimental evidence from radiology.” *NBER Working Paper* .
- Alufaisan, Yasmeen, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. “Does explainable artificial intelligence improve human decision-making?” *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (8):6618–6626.
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica* URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Baksi, Soham and Pinaki Bose. 2007. “Credence goods, efficient labelling policies, and regulatory enforcement.” *Environmental and Resource Economics* 37:411–430.
- Balafoutas, Loukas and Rudolf Kerschbamer. 2020. “Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions.” *Journal of Behavioral and Experimental Finance* 26:100285.
- Bao, Yongping, Ludwig Danwitz, Fabian Dvorak, Sebastian Fehrler, Lars Hornuf, Hsuan-Yu Lin, and Bettina von Helversen. 2022. “Similarity and consistency in algorithm-guided exploration.” *CESifo Working Paper* .
- Bigman, Yochanan E. and Kurt Gray. 2018. “People are averse to machines making moral decisions.” *Cognition* 181:21–34.
- Birnbaum, Michael H. and Ulrich Schmidt. 2015. “The impact of learning by thought on violations of independence and coalescing.” *Decision Analysis* 12 (3):144–152.
- Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann. 2019. “Task-dependent algorithm aversion.” *Journal of Marketing Research* 56 (5):809–825.

- 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.
- Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. “Algorithm aversion: people erroneously avoid algorithms after seeing them err.” *Journal of Experimental Psychology: General* 144 (1):114.
- . 2018. “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them.” *Management Science* 64 (3):1155–1170.
- Etilé, Fabrice and Sabrina Teyssier. 2016. “Signaling corporate social responsibility: Third-party certification versus brands.” *The Scandinavian Journal of Economics* 118 (3):397–432.
- European Commission. 2021. “Proposal for artificial intelligence act.” URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.
- Galton, Francis. 1907. “Vox populi.” *Nature* 75 (7):450–451.
- Green, Ben. 2022. “The flaws of policies requiring human oversight of government algorithms.” *Computer Law & Security Review* 45:105681.
- Green, Ben and Yiling Chen. 2019. “The principles and limits of algorithm-in-the-loop decision making.” *Proceedings of the ACM on Human-Computer Interaction* 3:1–24.
- Harbaugh, Rick, John W. Maxwell, and Beatrice Roussillon. 2011. “Label confusion: The Groucho effect of uncertain standards.” *Management Science* 57 (9):1512–1527.
- Hey, John D. 2001. “Does repetition improve consistency?” *Experimental Economics* 4:5–54.
- Jung, Markus and Mischa Seiter. 2021. “Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study.” *Journal of Management Control* 32 (4):495–516.

- Kim, Hyunjin, Edward L. Glaeser, Andrew Hillis, Scott D. Kominers, and Michael Luca. 2024. “Decision authority and the returns to algorithms.” *Strategic Management Journal* Forthcoming.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human decisions and machine predictions.” *The Quarterly Journal of Economics* 133 (1):237–293.
- Lai, Vivian, Chacha Chen, Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. “Towards a science of human-AI decision making: a survey of empirical studies.” *arXiv Working Paper* .
- Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. “Algorithm appreciation: People prefer algorithmic to human judgment.” *Organizational Behavior and Human Decision Processes* 151:90–103.
- Myagkov, Mikhail and Charles R. Plott. 1997. “Exchange economies and loss exposure: Experiments exploring prospect theory and competitive equilibria in market environments.” *The American Economic Review* 87 (5):801–828.
- Nicholls, Nicky, Aylit T. Romm, and Alexander Zimmer. 2015. “The impact of statistical learning on violations of the sure-thing principle.” *Journal of Risk and Uncertainty* 50:97–115.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting racial bias in an algorithm used to manage the health of populations.” *Science* 366 (6464):447–453.
- Önköl, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönöl, and Andrew Pollock. 2009. “The relative influence of advice from human experts and statistical methods on forecast adjustments.” *Journal of Behavioral Decision Making* 22 (4):390–409.
- Park, Joon S., Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. “A slow algorithm improves users’ assessments of the algorithm’s accuracy.” *Proceedings of the ACM on Human-Computer Interaction* 3:1–15.

- Pigors, Mark and Bettina Rockenbach. 2016. “Consumer social responsibility.” *Management Science* 62 (11):3123–3137.
- Prahl, Andrew and Lyn M. Van Swol. 2017. “Understanding algorithm aversion: When is advice from automation discounted?” *Journal of Forecasting* 36 (6):691–702.
- . 2021. “Out with the humans, in with the machines? investigating the behavioral and psychological effects of replacing human advisors with a machine.” *Human-Machine Communication* 2:209–234.
- Reich, Taly, Alex Kaju, and Sam J. Maglio. 2023. “How to overcome algorithm aversion: Learning from mistakes.” *Journal of Consumer Psychology* 33 (2):285–302.
- Sele, Daniela and Marina Chugunova. 2024. “Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making.” *Plos one* 19 (2):e0298037.
- van de Kuilen, Gijs. 2009. “Subjective probability weighting and the discovered preference hypothesis.” *Theory and Decision* 67:1–22.
- van de Kuilen, Gijs and Peter P. Wakker. 2006. “Learning in the Allais paradox.” *Journal of Risk and Uncertainty* 33:155–164.
- Yin, Ming, Vaughan Wortman, Jennifer Wortman, and Hanna Wallach. 2019. “Understanding the effect of accuracy on trust in machine learning models.” *Proceedings of CHI Conference on Human Factors in Computing Systems* :1–12.
- Zhang, Yunfeng, Vera Liao, and Rachel K. E. Bellamy. 2020. “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* :295–305.
- Zhang, Yunhao and Renee Gosline. 2022. “Understanding Algorithm Aversion: When Do People Abandon AI After Seeing It Err?” *SSRN Working Paper* .

Appendix A: Robustness Analysis: Excluding Unreasonably Small Guesses

As discussed in section 4.2.2, some of our participants state very low guesses, including 0. This section explores the effect of these small guesses on our overall results. We are particularly interested in the effect of the explanation and whether its impact on performance is driven by such very small guesses. We start our investigation by plotting pooled revised guesses from the two treatments with explanation against the revised guesses from the two treatments without explanation (figure 4.7). This figure does *not* include data from the VARYINGQUALITY condition, since that data was generated under alternating algorithm advice quality and can therefore not be compared to the other treatments. To focus on low values, figure 4.7 shows the revised guess range from 0 to 200.¹⁶

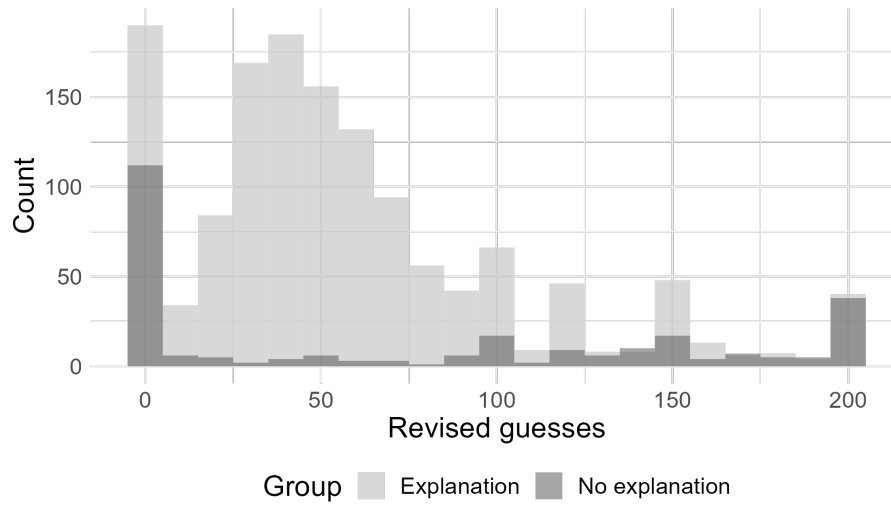
In figure 4.7 it is evident that both types of treatments (with and without explanation) exhibit a large number of guesses equal to 0. These 0 guesses are substantially more common in the explanation treatments. Moreover, again only for the explanation treatments, there is a separate smaller hump between 10 and 100 guesses.

How common are these low guesses? As the VARYINGQUALITY condition is not included in this graph, we have 1263 participants (instead of 1565 in the entire sample) who stated revised guesses in 16 rounds. We have therefore $1263 \times 16 = 20,208$ revised guesses. Out of these 1294 values, or 6,4% out of the 20,208, are below the value of 100, where 100 is an arbitrary, but not unreasonable threshold for serious guesses. Hence, the overwhelming majority of revised guesses are in a range above 100 guesses.

Nonetheless, one can ask if our main results are driven by these low values. In order to answer this question, one can examine figure 4.8, which shows the same information as

¹⁶For most of the paper, we analyze the natural logarithm as this is our way to address large outliers. In the following analyses, we are concretely interested in the small values and their interpretation and do not want to address them by taking the logarithm. We, therefore, analyze raw values.

Figure 4.7 Revised guesses between 0 and 200 pooled for treatments with and without explanation.



Notes: The figure shows raw values (unlogged).

our main result figure 4.3, except that all revised guesses below a value of 100 are excluded from the calculation.

One can see that our main results hold (with one small exception) when revised guesses below 100 are excluded from the analysis:

Result 1a: *Explanation reduces algorithm adherence (weakest effect).*

Result 1b: *Revealing the truth reduces algorithm adherence (medium effect).*

Result 1c: *Combining explanation and revealing truth reduces algorithm adherence (strongest effect).*

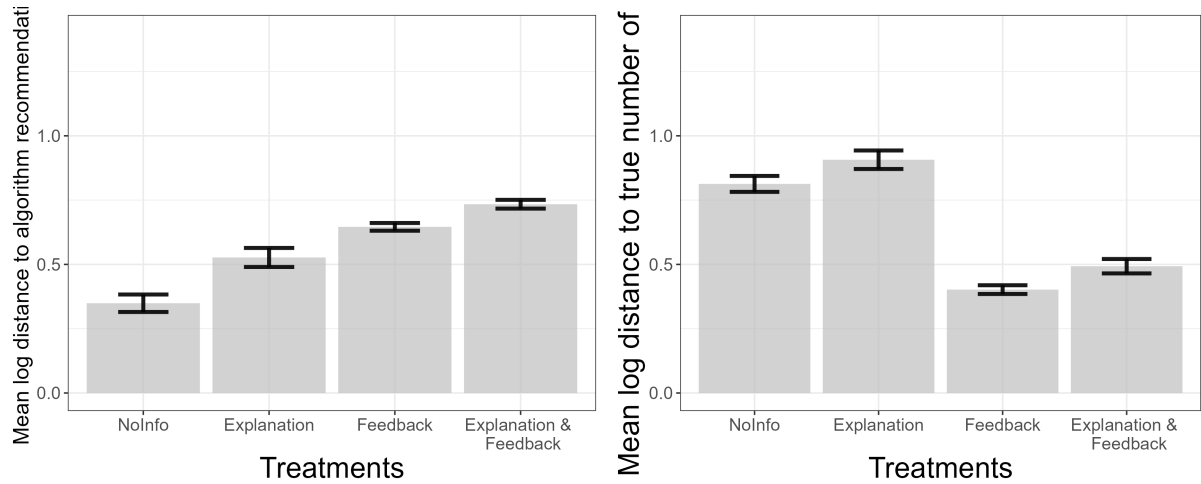
Result 2a: *Explanation does not improve performance (and possibly hurts).*

Result 2b: *Feedback improves performance.*

Result 2c: *When providing participants with both explanation and feedback, the net effect is equal to the positive and negative effects of the two individual treatments.*

The one change that appears is the sequence of effects on algorithm adherence. In the initial analysis, revealing the truth has the weakest effect on algorithm adherence. When

Figure 4.8 Mean distance to the algorithm and the true number of dots without guesses below 100



(a) Mean distance to the algorithm recommendation by treatment **(b)** Mean distance to the true number of dots by treatment

Notes: The bar graph in panel (a) illustrates the treatment effects on algorithm adherence (mean distance of the revised guesses to the algorithm recommendation per treatment). The bar graph in panel (b) illustrates the treatment effects on guessing performance (mean distance of the revised guesses to the true number of dots per treatment). The bar plots also include the standard errors around the mean. We pre-process the data by taking log values and calculating the mean over all 16 rounds for each individual. For more details see section 4.2.2.

values below 100 are excluded, revealing the truth has only the second weakest effect.

In the main results, explanation clearly hurts performance. In this robustness analysis, this negative effect is reduced. Still, there is a statistically significant difference as demonstrated by an independent two-sample t-test contrasting the mean outcomes of the baseline and explanation treatments, yielding a p-value of 0.047 ($t = -1.9844$, $df = 602.7$). Nonetheless, given that the robustness analysis markedly influences the effect size, we adopt a conservative stance and do not claim that explanation hurts performance. Our bottom line is that explanation does not improve performance, and there is a possibility that it may even hurt. The next section explores possible explanations as to why this might be the case.

Appendix B: Why Does Explanation Hurt Performance?

In the previous section, we have seen that very low guesses below 100 account for some part of the negative effect of explanation of performance. Figure 4.7 illustrates that only the treatments with explanation contain more guesses equal to zero and a hump shaped around 20-100. In this section, we address different hypotheses regarding what causes the 0-guesses and the humps in treatments with explanation.

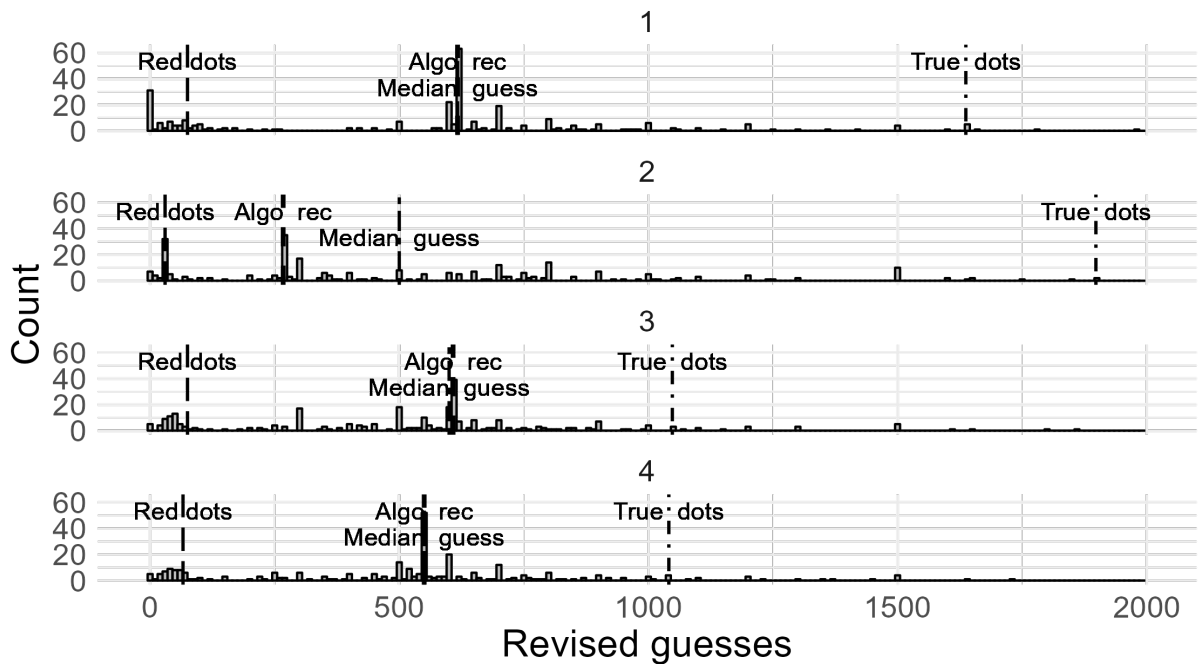
One possible explanation for the higher number of guesses equal to zero is that some of our participants were displeased or irritated by reading the explanation of how the algorithm works and consequently state a nonsensical guess. Contrary to this hypothesis, we observe that the number of revised guesses equal to zero decreases over the course of the experiment. This can be seen in figure 4.9, which shows a histogram of revised guesses in the EXPLANATION treatment for the first four rounds.

The trend of fewer zeros over the course of the rounds continues until the end of the experiment. Therefore, another possible explanation that is more in line with this observation is that some participants did not want to play this task. Note that our participants did not know that each round would contain the dot guessing task. Possibly, some were trying to move quickly on to a task they might find easier and stated a guess that required a low cognitive effort (namely a guess equal to zero).

We now turn to a discussion on what might have caused the hump shaped guesses around values of 20-100. One hypothesis is that participants who saw the explanation of how the algorithm worked interpreted this as an instruction. Instead of a critical assessment of the algorithm's quality, some participants might have concluded that the experimenter wants them to use the same approach as the algorithm to come up with their guess. Yet, if this were the case, we would expect to see more guesses around the algorithm recommendation - which is not the case - rather than in the range of 20-100.

A related hypothesis is that some participants misinterpreted our instructions and mis-

Figure 4.9 Histogram of revised guesses in the EXPLANATION treatment for the first four rounds

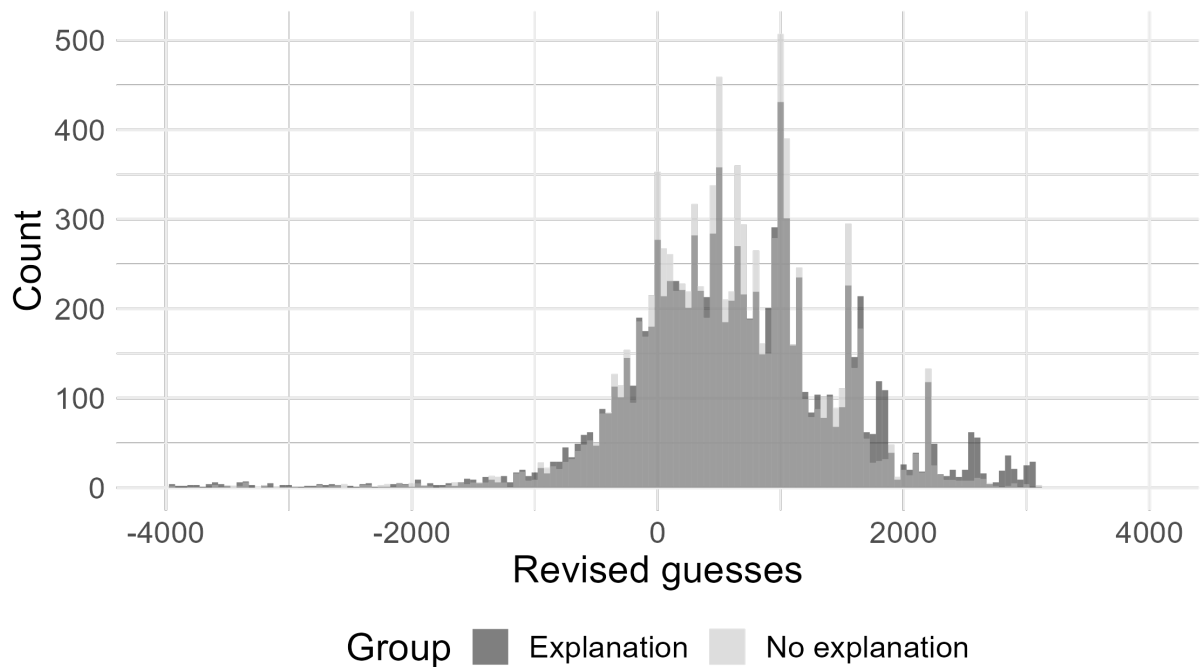


Notes: “True dots” indicates the true number of dots, “Median guess” indicates the median revised guess, “Algo rec” indicates the algorithm recommendation, and “Red dots” indicates the number of dots within the red squares of the explanation which participants in this treatment were exposed to. The figure shows raw values (unlogged).

takenly thought they should only guess the number of dots *within* the red boxes which they saw as the visual part of the explanation of how the algorithm works. But if this were the case, one would expect the hump shape to be centered above the number of dots within the red boxes. Figure 4.9 displays the guesses and the number of dots within the red boxes as a dotted dark red line. If this hypothesis was true, we would expect the number of dots in the red boxes to be in the center of the distribution humps. However, what we observe is that the hump shape is to the left of the dotted line in every round. If this was the driving behavior, participants must have systematically counted fewer dots than there actually were in the red boxes.

Another possibility is that some participants misinterpreted our instructions such that they should guess the *average* number of dots within *one* red box. This would be in line with the observed data but seems unlikely given that this would require them to read the instructions carefully enough to understand the algorithm, but at the same time to

Figure 4.10 Distribution of distances between revised guesses and true number of dots (non log values).



severely misinterpret the goal of the task, while making several incorrect assumptions.

Finally, we discuss the point that there is a statistically significant negative effect of explanation on performance even when low values are excluded. Figure 4.10 shows the distributions of differences between the revised guesses and the true number of dots (for non-log values), again pooled for the two treatments with and without explanation. From this, it becomes apparent that there is no obvious effect on the distribution that could drive the result. An eye-balling inspection could suggest that the distribution in the explanation treatments is more spread out. This could hint at a heterogeneous effect of explanation: Some participants might be adversely affected (due to boredom or other negative unintended externalities), while other participants correctly deduce that the algorithm advice is too low and over-correct for its bias. Yet, a Levene test for homogeneity of variances cannot reject the null of equal variances (F-value of 2.77 and a p-value of 0.095).

Overall, we cannot provide a definite answer to the question of why explanation hurts performance. Pinpointing the underlying mechanisms requires a new experimental design

and must remain an open question for future research. In this section, we have provided a discussion of the potential underlying mechanisms based on our experimental setup and the available data.

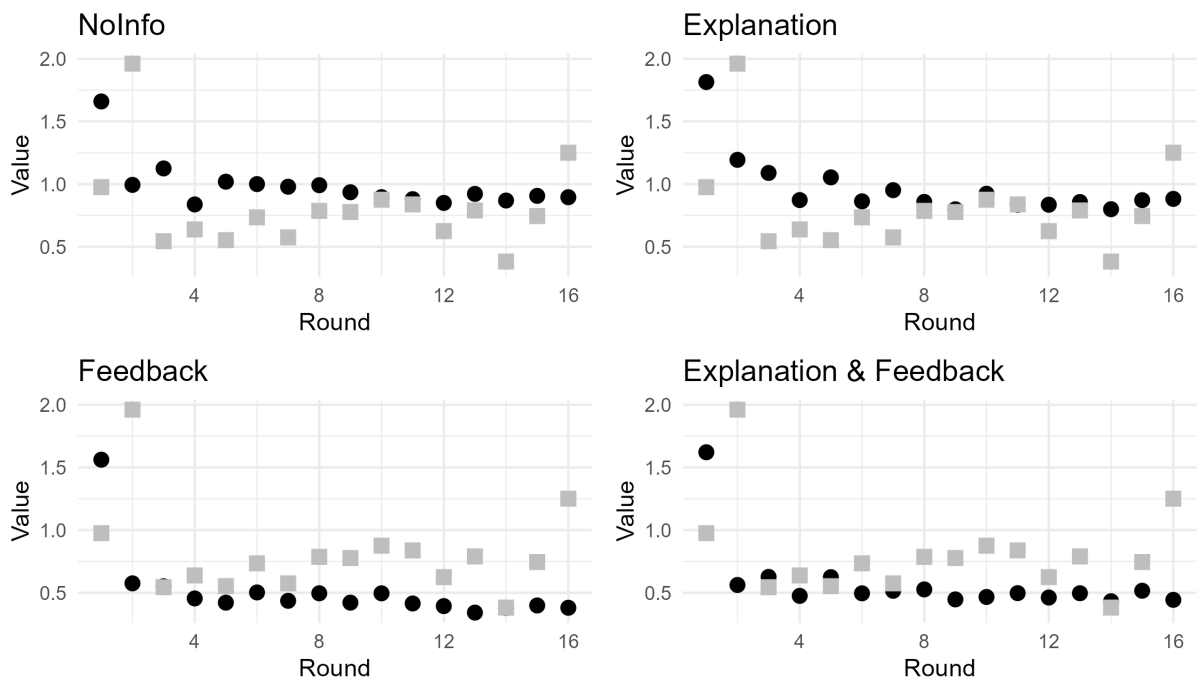
Appendix C: Would Participants Have Benefited from Following the Algorithmic Advice?

One relevant question is whether participants would have enhanced their performance by incorporating algorithmic advice into their revised guesses. We address this question for each round individually. The following figure illustrates the distance to the true answer for each round in every treatment. The squares denote the distance between the correct answer and the algorithmic recommendation, while the dots represent the average distance between the true values and the initial guess. Lower values indicate better performance. This figure allows to compare whether participants' *initial* guesses or the recommendation of the algorithm is closer to the true answer. On average, participants would have gained from assigning a positive weight to the algorithmic advice in most rounds in the baseline treatment (algorithmic prediction superior in 14/16 rounds) and the explanation treatment (algorithmic prediction superior in 13/16 rounds). Conversely, this is not observed in the feedback treatment (participant guesses superior in 14/16 rounds) and the combined treatment (participant guesses superior in 12/16 rounds).

Appendix D: Extreme Behaviors: Complete Adherence to Advice and Complete Disregard of Advice

The depicted figure illustrates the percentage of individuals who fully adhere to the algorithm (i.e., the revised guess matches the algorithmic prediction) and those who disregard the algorithmic advice entirely (i.e., the revised guess aligns with the initial guess). The

Figure 4.11 Performance participants vs. algorithm



Notes: The figure shows logged values.

figure demonstrates a consistent reduction in trust in algorithmic advice across all treatments. Further, treatments including feedback seem to elevate confidence in one's initial guess, likely due to the availability of an orientation point. In the explanation treatment, there is no apparent reason to put greater trust in one's initial guess compared to the baseline treatment. Consequently, in the explanation treatment, subjects exhibit diminished trust in the algorithm, but this doesn't translate into increased confidence in their own assessment.

Appendix E: Additional Analysis Baseline Treatment: Algorithmic Advice as a Credence Good

This section offers an additional interpretation of the baseline treatment, arguing that participants perceive the algorithmic advice, in our setting, as a credence good. However, we acknowledge that such a result is highly context-specific. Therefore, we don't put

Figure 4.12 Percentage of complete adherence to advice

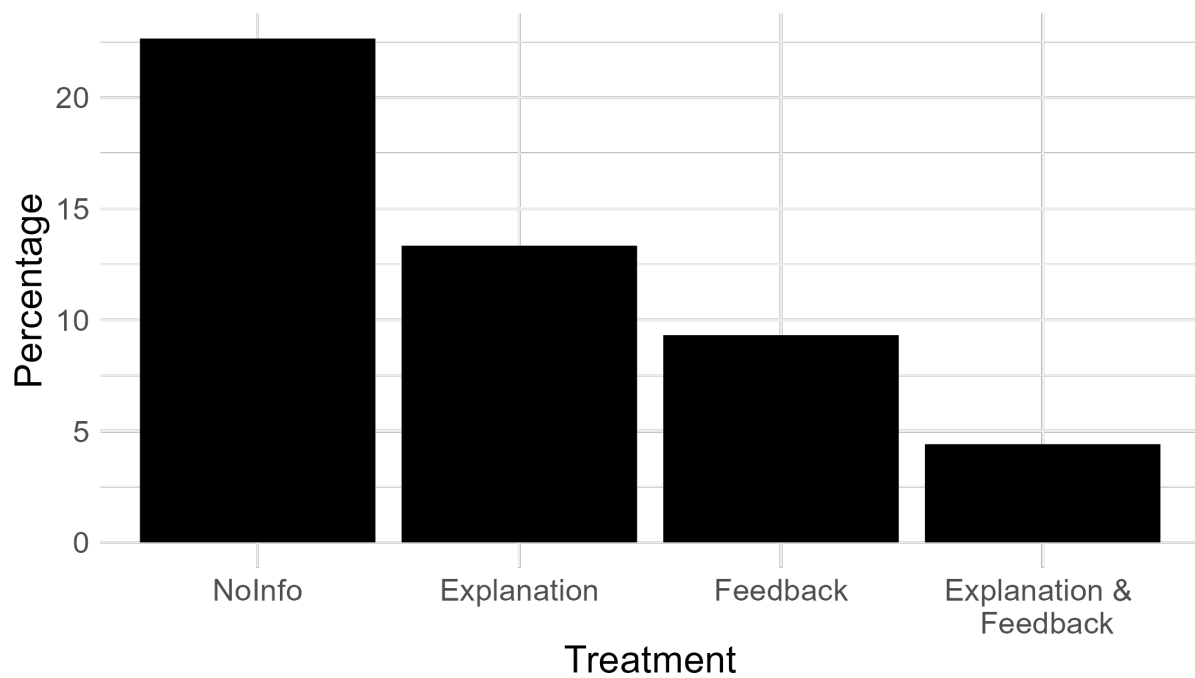
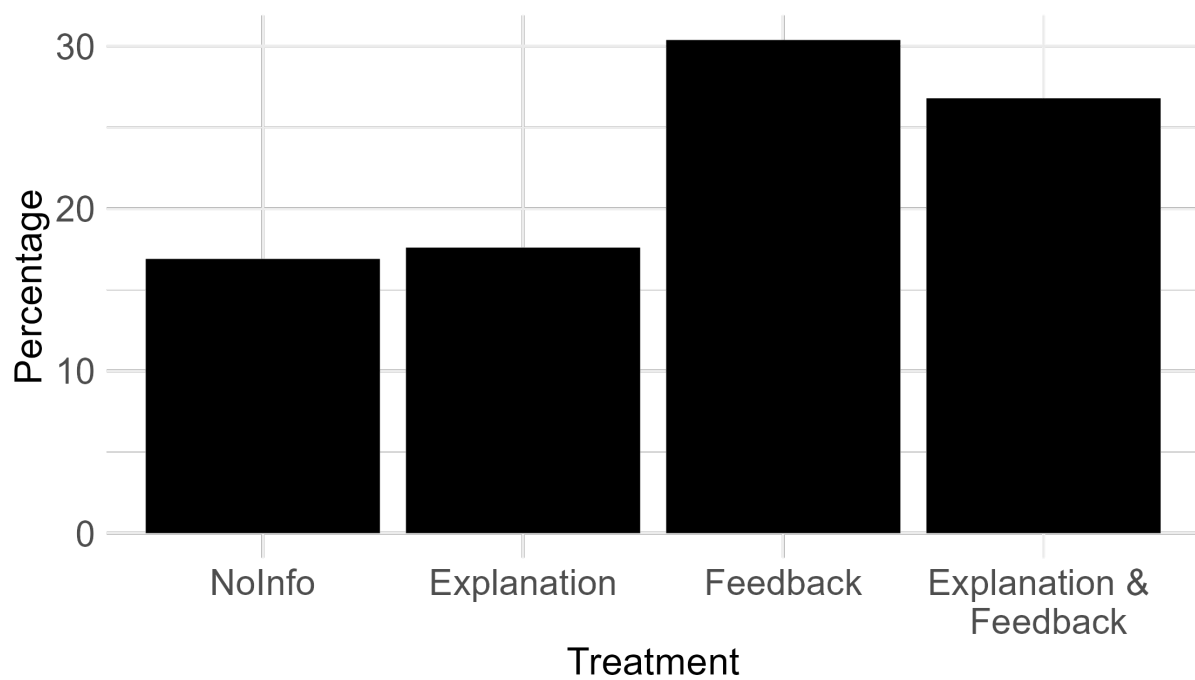


Figure 4.13 Percentage of complete disregard of advice



this interpretation forward in the main text. A common view among policymakers and in the academic literature implicitly assumes that human decision-makers can accurately assess the advice quality after observing the algorithm’s recommendation, for instance, by comparing it to their own judgment (e.g., in the Artificial Intelligence Act as proposed by the European Commission, 2021). From an economic point of view, this suggests that algorithm advice is perceived as an experience good, i.e., consumers can accurately assess the quality after consumption of the good. We challenge this assumption and argue that algorithms are often perceived to be *credence goods*. Even after “consumption” of the good – repeated interaction with the algorithm – humans cannot correctly assess its advice quality. Our baseline treatment provides experimental evidence from a reasonable setting that many people cannot correctly assess the quality of algorithmic advice even after “consuming” it. It follows that they perceive algorithmic advice as a credence good.¹⁷

Our work is hence related to the literature on credence goods that can be categorized into two different strands (Balafoutas and Kerschbamer, 2020). “Classical credence goods” involve asymmetric information between the expert seller and their customer regarding the fit between the characteristics of the products and the customer’s needs, prominent examples being healthcare or repair services. Our case aligns with the second strand of literature: “label credence goods.” Such goods have unobservable attributes that remain undetected after consumption, common examples being organically produced food and fairly traded products (Baksi and Bose, 2007; Harbaugh, Maxwell, and Roussillon, 2011; Pigors and Rockenbach, 2016; Etilé and Teyssier, 2016).

We ask the question of whether people can assess the quality of algorithmic advice (and if they act accordingly) after “consuming” this advice. To answer this question, one can inspect figure 4.14, which shows the densities of the initial and revised guesses of the first four rounds in treatment NOINFO.

In the first round, the initial guess density is flat: Participants vary vastly in their initial

¹⁷For our work, it is important that algorithms have unobservable attributes that remain undetected after consumption. This type of credence goods is often referred to as “label credence goods.” The analysis of “classical credence goods,” on the other hand, focuses on the asymmetric information between an expert seller and the customer. In this article, we refer to “label credence goods” whenever we employ the term “credence goods.”

guess. After observing the algorithm recommendation, participants state their revised guess, resulting in the revised guess density. Participants strongly react to the algorithmic advice and many subjects follow the advice closely. This can be directly inferred from the revised density: It is centered above the algorithm recommendation and its variance is greatly reduced. The second round illustrates that the initial guesses in round two are still influenced by the algorithmic advice from round one: The density of the initial distribution is centered above the previous round's algorithm recommendation. The algorithm recommendation in one round serves as an orientation point for the next initial guess. The revised guess density in round two peaks again above the algorithm recommendation. In rounds three and four one again observes the two effects (1) the revised guesses move closer to the algorithm and (2) their variance is reduced (both compared to the initial guess density in the same round). In fact, this pattern holds for all subsequent 12 rounds (see figures 4.15 to 4.18).

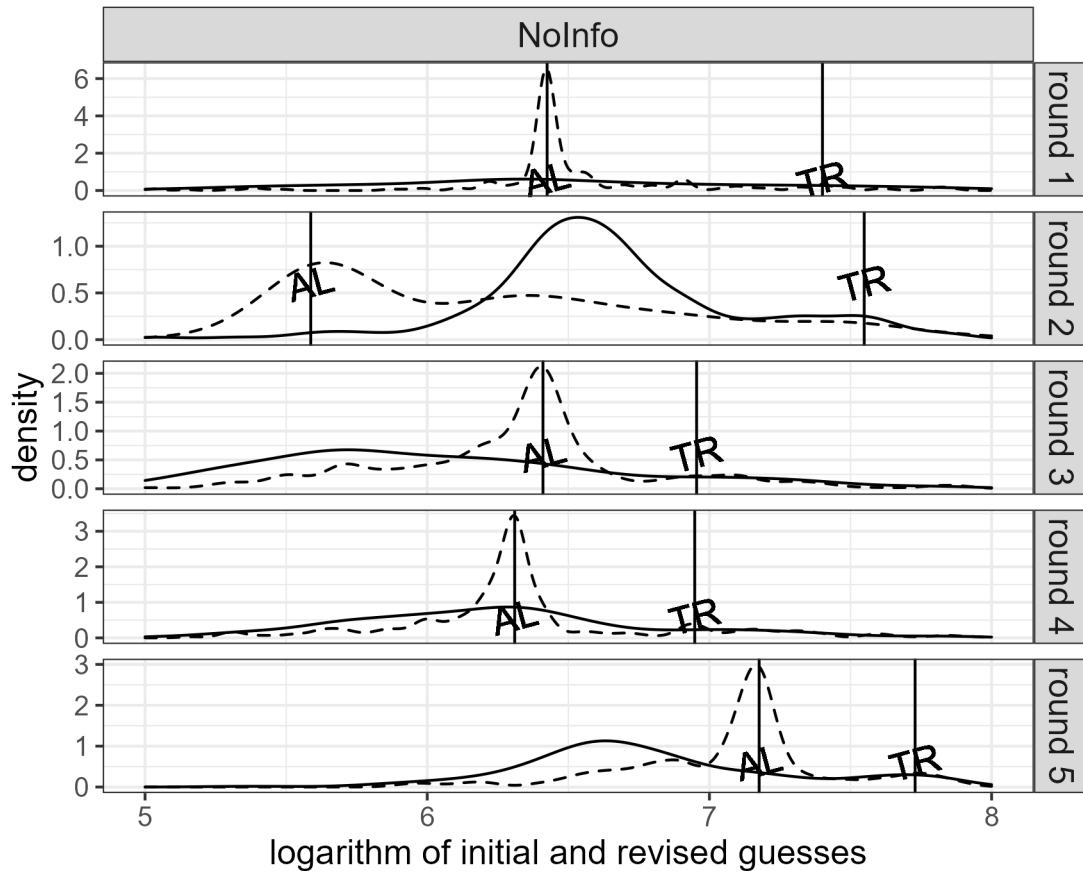
Table 4.1 quantifies these differences and shows the average of the individual log distances to the algorithm and the standard deviation for the initial and revised guesses. It also exhibits p-values from a t-test comparing the initial and revised distance to the algorithm recommendation and Levene's test for homogeneity of variances of the initial and revised densities.

The average distance to the algorithm recommendation is smaller for the revised guesses than for the initial guesses in all 16 rounds (this difference is always significant except for one round). In other words, revised guesses move closer to the algorithm. The standard deviations of the revised guess densities are smaller for the revised guesses in a majority of cases.¹⁸

If the algorithmic advice exhibited traits of an experience good, we would expect our subjects to learn to optimally incorporate this advice into their decision-making and adjust how strongly they adhere to the algorithm. Note that in each round, participants have the possibility to compare the algorithm recommendation with their own guess, which was elicited in the first stage. Given their own guess as a reference point, participants

¹⁸Note that many of the p-values of these differences are not significant.

Figure 4.14 Densities of initial and revised guesses for the first five rounds of NOINFO



Notes: Initial (black line) and revised (dotted line) guess densities for NOINFO for the first five rounds. Algorithm recommendation (AR) is the leftmost vertical line and the true number of dots (TR) is indicated by the rightmost vertical line. The figure only shows the range of $\log(\text{guess})$ from 5 to 8 and therefore does not display the tails of the distributions. The range of the axis showing the density differs between rounds.

could realize that the algorithm recommendation is consistently too low. Over time (i.e., over rounds), if participants had this realization repeatedly, and assuming they would also optimally react based on this insight, we would see a shift in the revised guess density away from the algorithm (and potentially closer to the true number of dots). As can be seen from figure 4.14 and table 4.1 there is no evidence for this. In figure 4.14 there is no increase in probability mass in the region above the algorithm recommendation for the revised guesses over rounds. In table 4.1 average distance of the revised guesses to the algorithm is in every round smaller than the distance for the initial guesses. In other words, the average revised guess always move closer to the algorithm. So the participants never learn to move further away from the algorithm recommendation (or ignore it).

Table 4.1 Distances to algorithmic recommendation per round: NOINFO

Round	$ \log(algo) - \log(guess_i) $		$ \log(algo) - \log(guess_r) $		p-values	
	mean	sd	mean	sd	t-test	levene
1	1.29	1.78	0.59	1.39	0.00	0.00
2	1.20	0.73	0.88	1.13	0.00	0.00
3	0.78	0.93	0.42	0.74	0.00	0.13
4	0.51	0.67	0.45	1.02	0.28	0.33
5	0.66	0.89	0.35	0.74	0.00	0.77
6	0.67	1.16	0.42	1.00	0.00	0.22
7	0.60	1.06	0.33	0.90	0.00	0.06
8	0.61	0.91	0.37	0.81	0.00	0.21
9	0.58	0.96	0.42	1.03	0.00	0.66
10	0.62	0.99	0.40	0.98	0.00	0.60
11	0.46	0.74	0.31	0.90	0.00	0.69
12	0.47	0.70	0.29	0.68	0.00	0.68
13	0.53	1.08	0.33	0.89	0.00	0.19
14	0.64	0.74	0.42	0.95	0.00	0.82
15	0.52	1.07	0.34	1.01	0.00	0.31
16	0.68	0.89	0.49	0.98	0.00	0.24

Notes: Table contains the mean and standard deviation of the distance between the guesses and the algorithmic recommendation for every round. This distance is included with respect to the initial and the revised guesses. All values are logs. The table also contains the p-values for the t-test (difference between means) and the Levene-test (differences between variances) to test the difference between the values for initial and revised guesses for every round. Values refer to NOINFO.

Especially participants whose initial guesses fall within this range between the algorithm and the true value should not move closer to the algorithm to maximize their payoff. Yet, our data show that participants move closer to the algorithmic advice. Our results therefore rather indicate that the algorithmic advice exhibits traits of a credence good: Even after “consuming” the advice repeatedly, our participants appear not to be able to assess the quality of the advising algorithm.

Consequently, we document the following result:

Result: *The human decision-makers perceive the algorithmic advice as a credence good.*

Participants in the baseline treatment NOINFO repeatedly see biased advice in a task that does not require expert knowledge. Akin to many real-life decision situations, in this treatment, they therefore can compare their own prediction (note that we prime participants by specifically eliciting their initial guesses before seeing the algorithm) with the algorithm prediction. In principle, this could allow any participant to realize that the algorithm is biased and produces dot predictions that are too low. The question is how many participants realize this and if they can correctly adjust for this bias.

From the written feedback that participants could state at the end of the experiment, we know that some participants in the baseline treatment indeed recognize that the algorithm is downward biased (e.g., “I thought the algorithm consistently underestimated the number of dots”, “I did not trust the algorithm. It seemed to be generating numbers that were too low.”), yet others fail to assess the existence, magnitude, or direction of the bias (e.g., “I quickly began to depend on the algorithm and as the study progressed, got close to guessing what the algorithm predicted”, “I figured that it was overestimating”). Ultimately, the question about the nature of the algorithm as a good is an empirical one: How do the majority of participants assess the algorithm? As described previously, most participants follow the algorithm closely and never realize that they could improve their payoff if they move away from a biased recommendation.

Some important caveats are in order. We do not claim that all algorithms are always perceived as credence goods. We merely point out that algorithms can be perceived as credence goods in many applications. Clearly, one important determinant is how the human and algorithm abilities compare. Our results and implications could be considered in scenarios where neither human nor algorithm is obviously better, but where there is ambiguity about performance comparison. Moreover, the nature of our task is rather mathematical and objective, and behavior may vary for more subjective tasks (Castelo, Bos, and Lehmann, 2019). Further, our participants may assume that the algorithm performs better than humans because it would not be employed otherwise. In sum, one might consider subjects' priors that the algorithm is more capable to be justified. In addition, one might criticize that learning is generally not possible without feedback and explanation.¹⁹

In this section, we have stated the general assumption that individual humans can successfully provide oversight for algorithms is flawed and discuss the implications of this. A natural adjustment is, therefore, to put less emphasis on *individual* oversight and instead shift the focus to *collective* oversight. For example, organizations could audit algorithmic advice systems. This could entail checking possible training data, systematically challenging the algorithm, or controlled human-subject field experiments before deployment.²⁰

Our results have implications for both policymakers and managers. We show that there are situations in which humans are not able to accurately assess an advising algorithm's quality. In the context of regulation, this casts doubt on the effectiveness of individual human decision-makers to recognize biased algorithm predictions and to correct this bias to prevent harm. In the context of management, it means that organizations generally can neither rely on individual decision-makers to optimize decisions, nor can they rely on feedback from their decision-makers about the quality of an algorithm as a product.

¹⁹In response, we argue that what makes learning possible is the repeated engagement with the task and the fact that subjects provide initial as well as revised guesses.

²⁰This idea is similar to Green (2022), who suggests "institutional oversight."

Appendix F: Tables and Figures

Table 4.2 Treatment effect on algorithm adherence: Log-distance to algorithm recommendation: Overview

Treatment	n	min	mean	max	std. err.
NoInfo	324	0.001	0.426	9.514	0.035
Explanation	314	0.001	0.842	5.188	0.052
Feedback	312	0.001	0.711	1.989	0.017
Explanation & Feedback	313	0.001	0.969	4.271	0.034

Notes: The values in this table refer to the barplot in panel (a) of figure 4.3a. “Std. err.” refers to the standard error of the mean.

Table 4.3 Treatment effect on guessing performance: Log-distance to true number of dots: Overview

Treatment	n	min	mean	max	std. err.
NoInfo	324	0.001	0.897	9.122	0.033
Explanation	314	0.001	1.258	4.678	0.057
Feedback	312	0.001	0.479	2.025	0.022
Explanation & Feedback	313	0.001	0.792	5.087	0.052

Notes: The values in this table refer to the barplot in panel (b) in figure 4.3. “Std. err.” refers to the standard error of the mean.

Table 4.4 Distances to algorithmic recommendation per round: EXPLANATION

	$ \log(algo) - \log(guess_i) $		$ \log(algo) - \log(guess_r) $		p-values	
Round	mean	sd	mean	sd	t-test	levne

1	1.45	1.95	1.26	1.89	0.14	0.58
2	1.44	1.19	1.24	1.16	0.00	0.04
3	0.82	1.12	0.84	1.17	0.77	0.04
4	0.65	0.93	0.79	1.14	0.03	0.00
5	0.75	0.94	0.91	1.46	0.04	0.00
6	0.66	0.99	0.80	1.25	0.06	0.00
7	0.63	0.98	0.76	1.30	0.05	0.00
8	0.62	0.77	0.70	1.02	0.19	0.00
9	0.60	0.69	0.78	1.20	0.00	0.00
10	0.70	1.02	0.79	1.12	0.20	0.00
11	0.58	0.68	0.75	1.13	0.00	0.00
12	0.56	0.82	0.75	1.21	0.00	0.00
13	0.60	0.95	0.73	1.12	0.04	0.00
14	0.63	0.66	0.69	1.02	0.34	0.00
15	0.59	0.87	0.82	1.33	0.00	0.00
16	0.81	0.81	0.86	1.01	0.32	0.00

Notes: The structure of the table is the same as in table 4.1, but the values refer to EXPLANATION.

Table 4.5 Distances to algorithmic recommendation per round: FEEDBACK

Round	$ \log(algo) - \log(guess_i) $		$ \log(algo) - \log(guess_r) $		p-values	
	mean	sd	mean	sd	t-test	levene
1	1.25	1.76	0.44	1.05	0.00	0.00
2	1.71	0.72	1.52	0.94	0.00	0.00
3	0.71	0.86	0.68	0.87	0.58	0.50

4	0.80	0.87	0.72	0.75	0.14	0.73
5	0.52	0.35	0.59	0.99	0.23	0.01
6	0.82	0.89	0.77	0.82	0.20	0.94
7	0.56	0.66	0.51	0.64	0.36	0.99
8	0.76	0.74	0.63	0.52	0.00	0.15
9	0.85	0.75	0.71	0.42	0.00	0.13
10	0.86	0.77	0.75	0.71	0.06	0.93
11	0.75	0.60	0.69	0.51	0.11	0.74
12	0.64	0.70	0.61	0.69	0.47	0.86
13	0.76	0.58	0.70	0.60	0.06	0.58
14	0.42	0.37	0.45	0.69	0.39	0.11
15	0.65	0.51	0.59	0.50	0.00	0.70
16	1.07	0.54	1.01	0.56	0.12	0.61

Notes: The structure of the table is the same as in table 4.1, but the values refer to FEEDBACK.

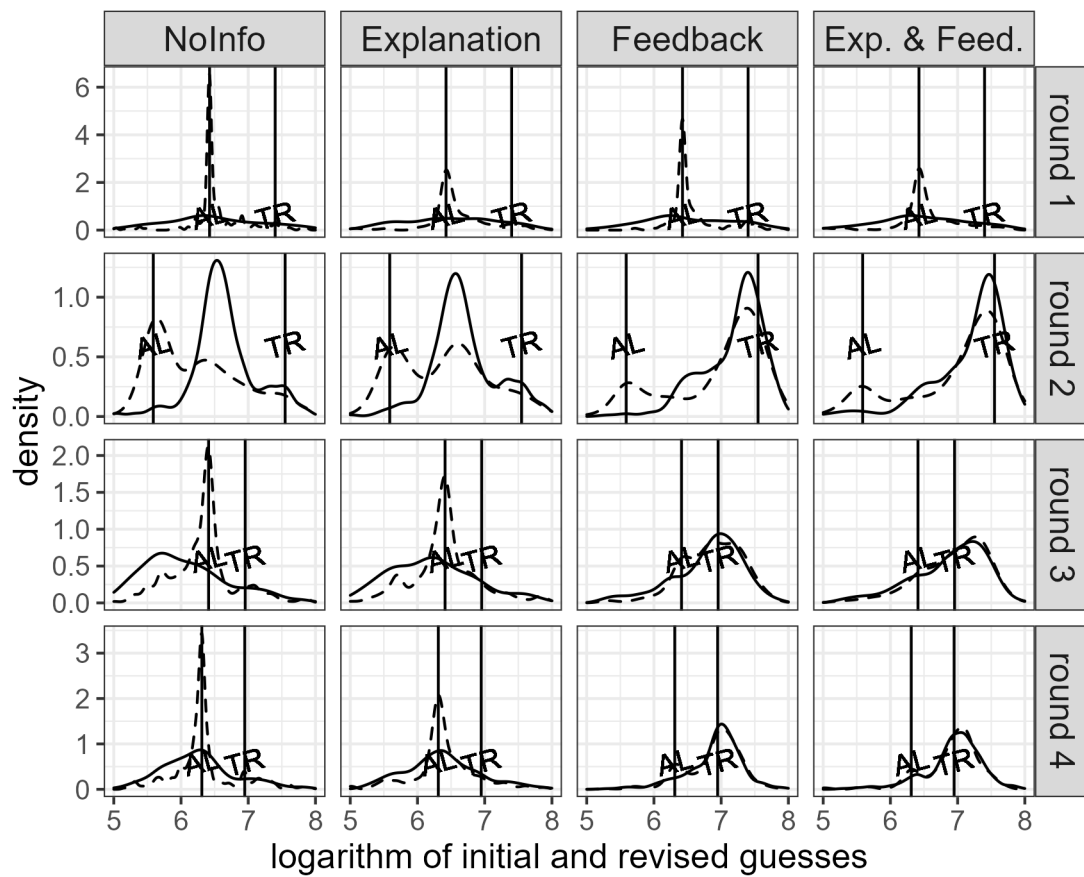
Table 4.6 Distances to algorithmic recommendation per round: EXPLANATION&FEEDBACK

Round	$ \log(algo) - \log(guess_i) $		$ \log(algo) - \log(guess_r) $		p-values	
	mean	sd	mean	sd	t-test	levene
1	1.23	1.86	1.25	1.89	0.90	0.14
2	1.74	0.71	1.64	0.96	0.09	0.00
3	0.81	0.95	0.92	0.99	0.06	0.19
4	0.84	0.82	0.91	0.90	0.19	0.19
5	0.67	1.02	0.89	1.35	0.00	0.01

6	0.85	0.84	0.95	0.91	0.07	0.12
7	0.64	0.85	0.74	0.91	0.15	0.12
8	0.79	0.87	0.91	1.05	0.06	0.08
9	0.89	0.79	0.96	0.85	0.15	0.16
10	0.89	0.71	1.07	1.26	0.01	0.00
11	0.86	0.79	0.93	0.92	0.16	0.10
12	0.73	0.82	0.85	0.89	0.04	0.16
13	0.88	0.90	0.91	0.88	0.61	0.48
14	0.51	0.67	0.59	0.70	0.08	0.11
15	0.75	0.80	0.84	0.90	0.12	0.05
16	1.10	0.63	1.14	0.68	0.41	0.04

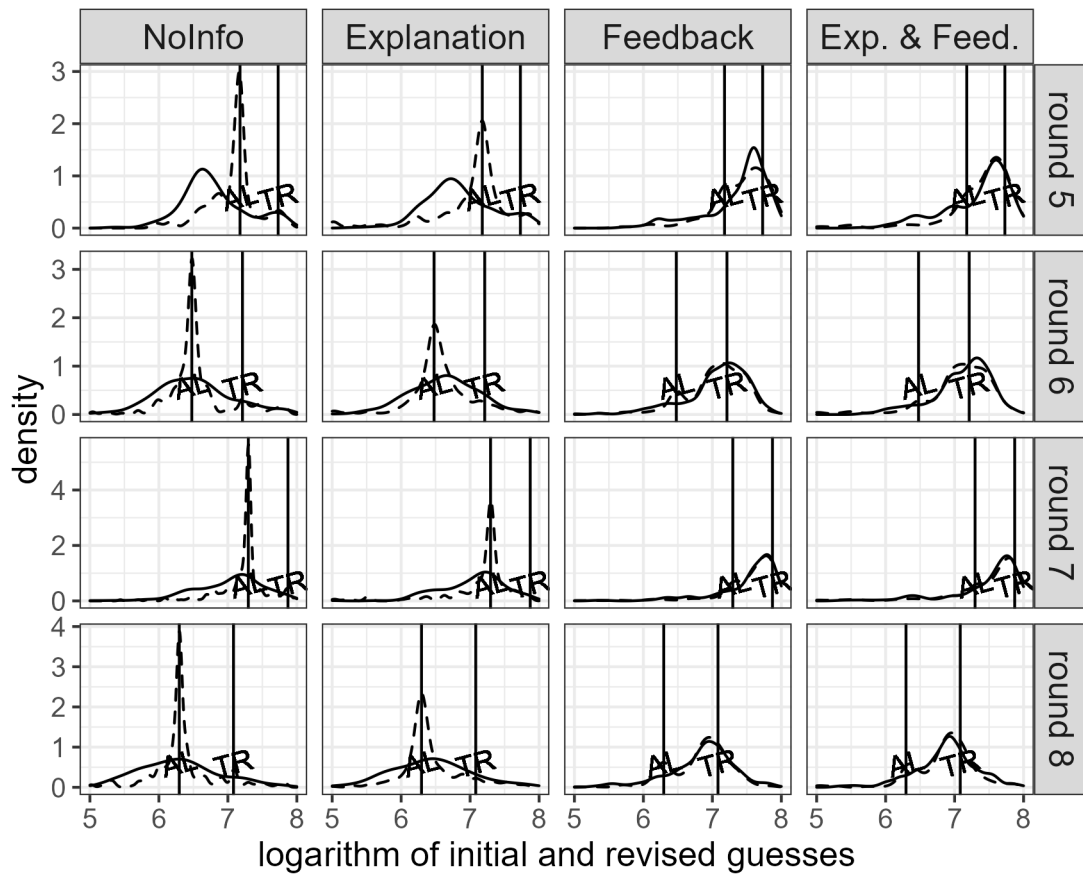
Notes: The structure of the table is the same as in table 4.1, but the values refer to EXPLANATION&FEEDBACK.

Figure 4.15 Distribution of initial and revised guesses by treatment for rounds 1 to 4



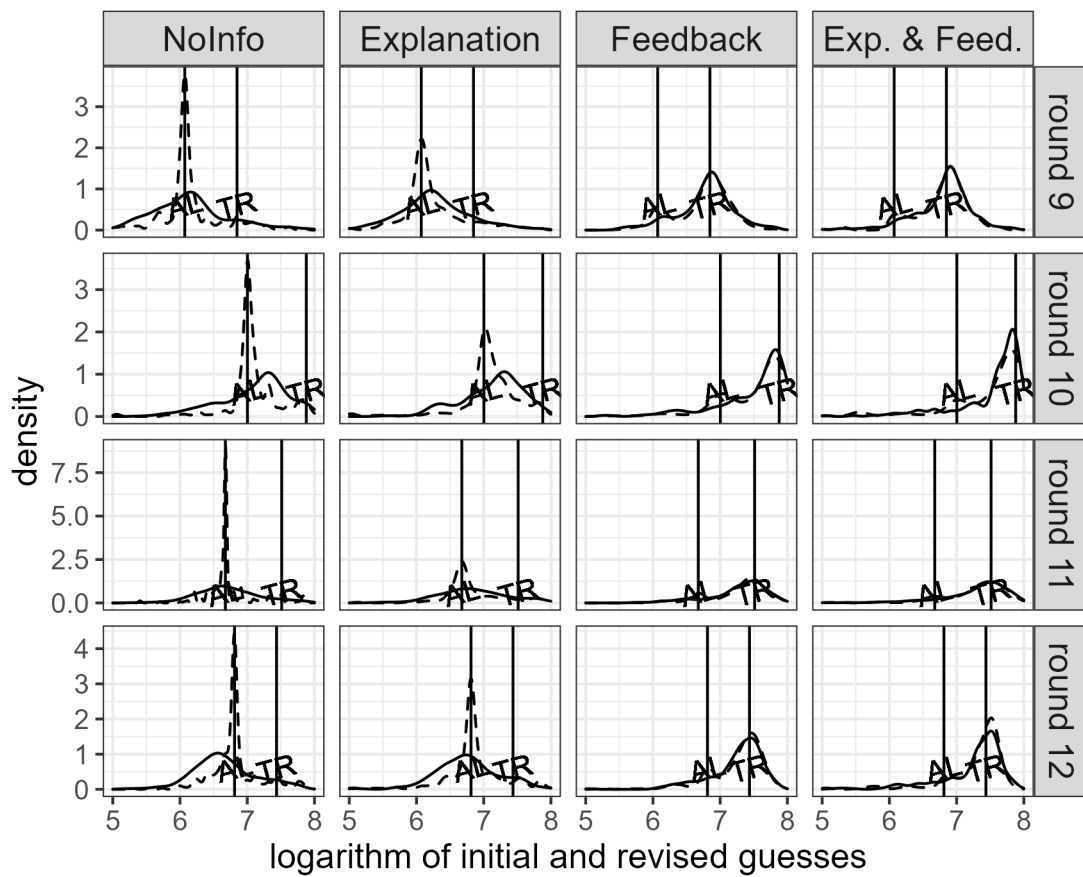
Notes: Initial and revised guess densities for round 1 to 4.

Figure 4.16 Distribution of initial and revised guesses by treatment for rounds 5 to 8



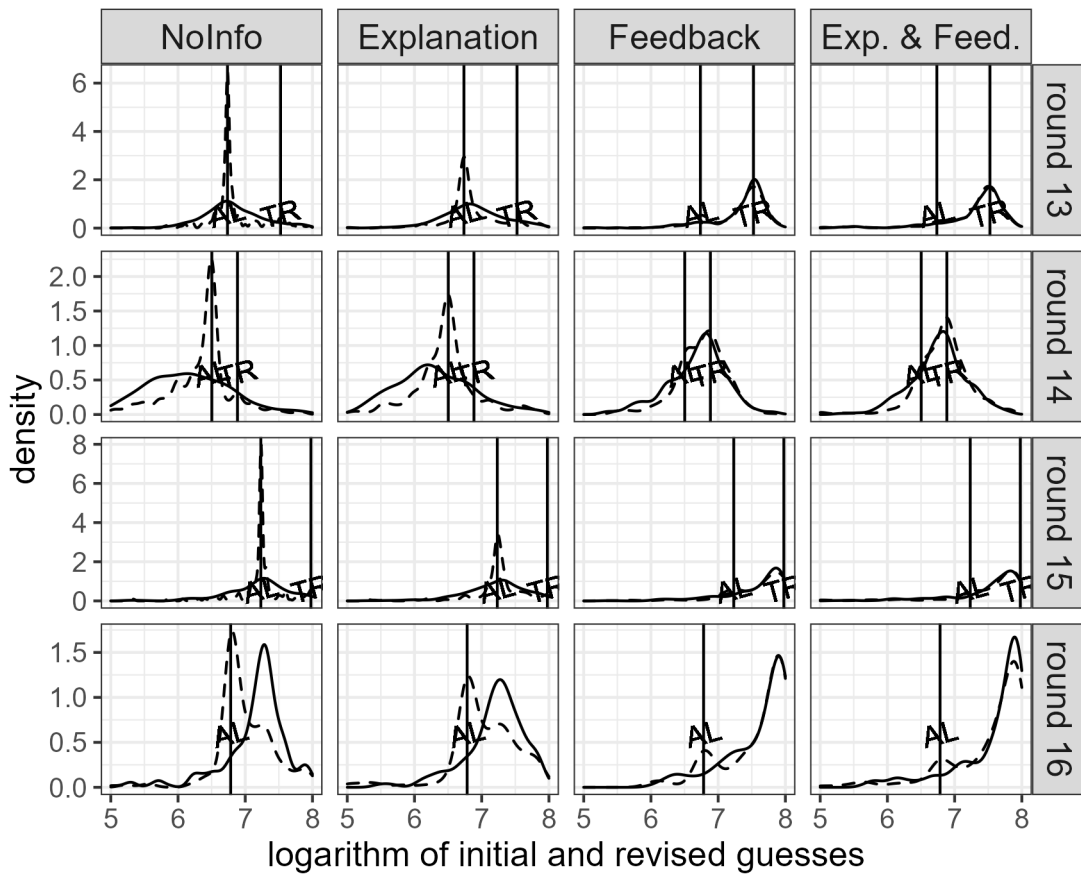
Notes: Initial and revised guess densities for round 5 to 8.

Figure 4.17 Distribution of initial and revised guesses by treatment for rounds 9 to 12



Notes: Initial and revised guess densities for round 9 to 12.

Figure 4.18 Distribution of initial and revised guesses by treatment for rounds 13 to 16



Notes: Initial and revised guess densities for round 13 to 16.

Appendix G: Experimental Interfaces

Figure 4.19

Welcome & Informed Consent

Thank you for participating in this study! The purpose of this study is to explore human decision-making.
This study is anonymous. We will not ask for your name or any information that will make you identifiable.
There is **no deception** in this study. Everything you see or read is true.

The study takes most participants less than 10 minutes to complete.
You will receive a fixed payment of \$0.90 (base reward) for your participation. You will also have the chance to earn up to \$4.80 additional dollars depending on your behavior during the study (bonus rewards).

The risks to participating are no greater than those encountered in everyday life. Your participation in this study is completely voluntary, and you may refuse to participate or withdraw from the study without penalty. Compensation will be awarded upon completion of the entire study. If you have any questions, please contact us via MTurk. Please feel free to print or save a copy of this consent form.

Please tick the following box to be able to continue:
 I have read and understood this consent form and wish to participate in this study.

Next

Notes: All participants saw this text as their first page.

Figure 4.20

Feedback

You're almost done!

We value your feedback! Did you find anything unclear or misleading? Any technical issues? Any other feedback regarding any aspects of the study? Would you like to explain your behavior in the study? (For example, did you trust the algorithm? How did this develop over time?) Let us know!

Feedback

Click below to receive the completion code and finish the study.

Next

Notes: All participants saw this as their final page before exiting the experiment.

Figure 4.21

Final Result

Your bonus payoff is \$0.00

You have completed the study. Your completion code is MERRY_CHRISTMAS. Please copy this code and return to MTurk.

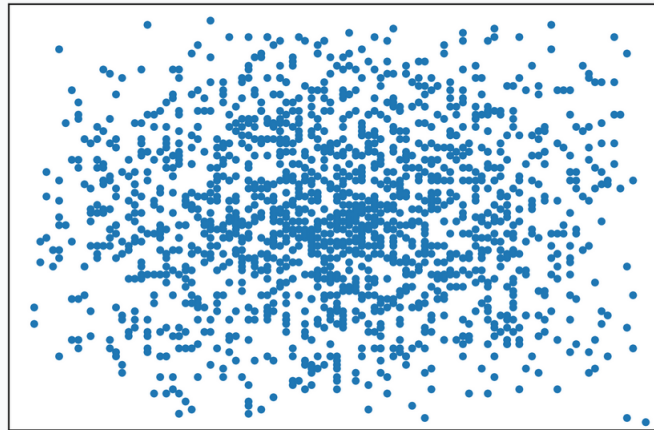
Notes: All participants learned about their final payoff and received a completion code.

Figure 4.22

Guess: How many dots are in this graph? Round 1/2

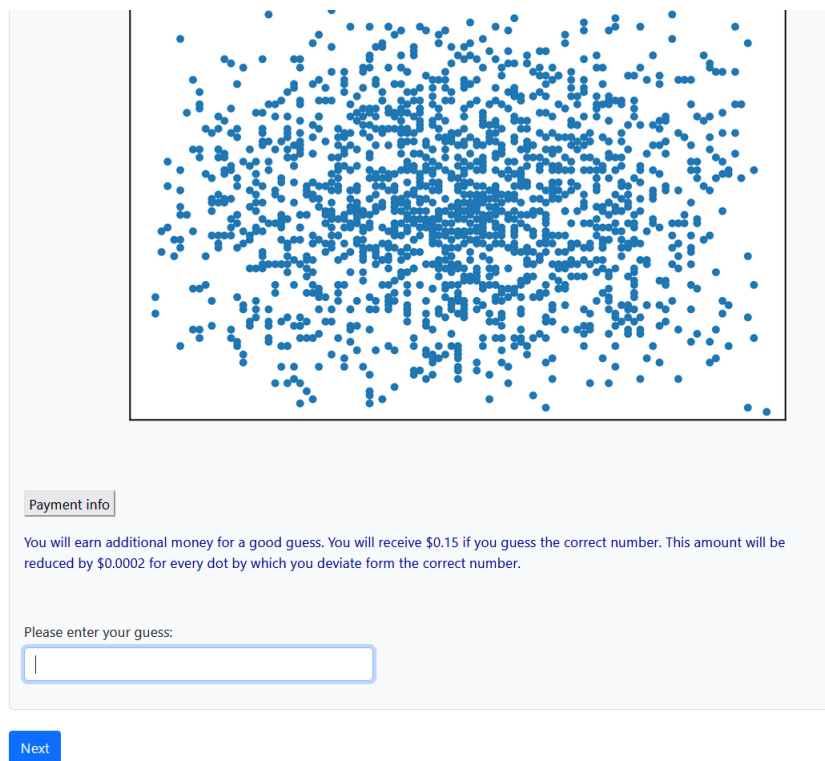
Time left to complete this page: 0:49

Below you see a new image.



Notes: Upper part of initial guess page: All participants saw this page (with the number of dots changing from round to round) before stating their initial guess. Participants had 60 seconds to state their guess.

Figure 4.23



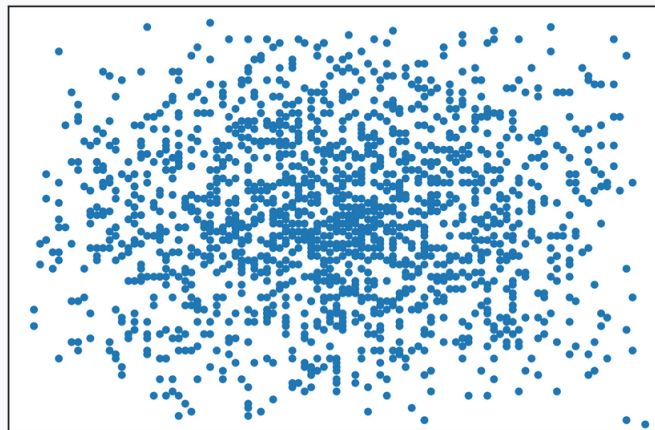
Notes: Lower part of initial guess page: All participants saw this when stating their initial guess.

Figure 4.24

Machine guess: Round 1/2

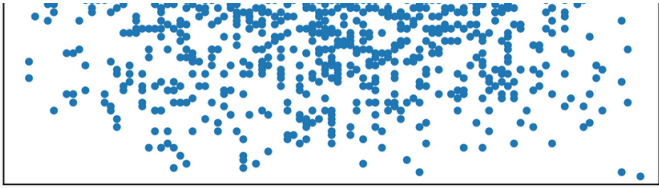
Time left to complete this page: 0:23

Below you see the image from the [previous](#) page.



Notes: In treatments without explanation, participants saw the same image again before stating their revised guess. In

Figure 4.25



We will now provide you with an algorithmic prediction regarding the number of dots in this picture. This prediction comes from an algorithm that is trying to solve the same problem as you have.

The algorithm predicts that there are 617 dots in this picture.

Your initial guess was 2333.

In light of the new information, you can now modify your guess. Please enter your revised guess below.

Payment info

You will earn additional money for a good guess. You will receive \$0.15 if you guess the correct number. This amount will be reduced by \$0.0002 for every dot by which you deviate from the correct number.

Please enter your revised guess:

Next

Notes: All participants saw the algorithm prediction before stating their revised guess.

Figure 4.26

Result Round 1/2

Your initial guess was 2333.

Your final guess was 333.

This round has ended. Click below to get to the next round.

Next

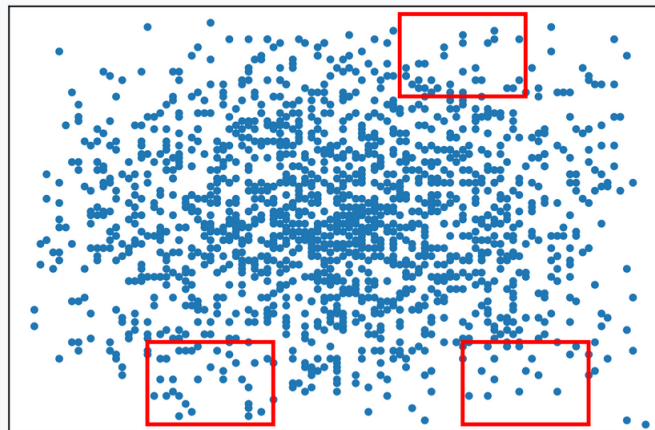
Notes: Participants in treatments in which the true answer was not revealed saw their initial and revised guess again, before moving on to the next round.

Figure 4.27

Machine guess: Round 1/2

Time left to complete this page: 0:51

Below you see the image from the [previous](#) page.



Notes: Participants in treatments in which the algorithm was explained saw the same image again but overlaid with red rectangles indicating the areas from which the algorithm sampled.

Figure 4.28

We will now provide you with an algorithmic prediction regarding the number of dots in this picture. This prediction comes from an algorithm that is trying to solve the same problem as you have.

The algorithm predicts that there are 617 dots in this picture.

Explanation of algorithm:

1. The algorithm counts the number of dots within each red square.
2. It then calculates the average of dots over the three squares.
3. Finally, it multiplies this average by 25, because 25 squares cover the entire area.

Your initial guess was 33.

In light of the new information, you can now modify your guess. Please enter your revised guess below.

Payment info

You will earn additional money for a good guess. You will receive \$0.15 if you guess the correct number. This amount will be reduced by \$0.0002 for every dot by which you deviate from the correct number.

Please enter your revised guess:

Notes: Participants in treatments in which the algorithm was explained additionally were provided a verbal explanation of the algorithm.

Figure 4.29

Result Round 1/2

Your initial guess was 22.

Your final guess was 33.

The true number was 1637.

This round has ended. Click below to get to the next round.

Next

Notes: Participants in treatments in which the true answer was revealed saw their initial and revised guess as well as the solution before moving on to the next round.

In the VARYINGQUALITY treatment the user interface looked identical to the EXPLANATION&FEEDBACK treatment, except for the dot distribution, which alternated every round between being uniformly and triangularly distributed. An example of a uniform dot distribution can be seen in figure 4.5.

Anhang der Dissertation

Liste der aus dieser Dissertation hervorgegangenen Veröffentlichungen

Chapter 2:

Aktuelle Version: nicht publiziert.

Eine ältere Version wurde als SSRN Working Paper unter dem Titel “Does Discussing How Much to Share Affect Sharing Behavior and Trust?” veröffentlicht:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3945496

Chapter 3:

Nicht publiziert.

Chapter 4:

Aktuelle Version: nicht publiziert.

Eine ältere Version wurde als SSRN Working Paper unter dem Titel “Algorithmic Advice as a Credence Good” veröffentlicht:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3945496

Abstract

This dissertation contains three empirical essays. The first essay examines the effects of a debating process prior to a collective decision. We conduct a lab-in-the-field experiment with school minors in Germany. We randomly assign some of them to discuss via chat how much they want to donate to a charity supporting incoming refugee minors. In our study, the pre-vote debate leads to higher donations. It does not directly affect trust among discussants, but subjects who are perceived as refugee-friendly after the discussion benefit from increased trust from their chat partners.

The second essay addresses belief formation and asks how breaking a social norm affects what people think others would do in the same situation. To study this question, I employ an online experiment with a representative sample of the UK's general population exposing subjects to either a high or a low temptation to lie. I find that subjects in the tempting environment lie more and, importantly, hold more pessimistic beliefs about what others would do in the same situation. I also include additional treatments in order to argue that the observed effects are not driven by rational expectations. The study provides causal evidence that breaking a social norm leads to strategic belief distortion about other people's behavior in order to maintain a positive self-image.

The third essay studies a situation in which a human decision-maker is assisted by an algorithm. We conduct an online experiment with US participants who repeatedly perform an estimation task while receiving (largely biased) recommendations from an algorithm. We analyze two interventions and ask whether they can help humans to assess the quality of algorithmic advice. First, we find that explaining the functioning of the algorithm in abstract terms reduces adherence to algorithmic advice, but it does not improve decision-making performance. Second, disclosing the correct answer after each round reduces adherence to algorithmic advice and improves human decision-making performance. While existing literature suggests that people abandon algorithms after seeing them err, this is not confirmed in our setting. This is likely because in our setting people can comprehend

the reasons why some of the algorithmic predictions are inaccurate. Jointly, the three essays provide insights into (behavioral) economic concepts of trust, generosity, belief formation, and advice taking through the lens of experimental methods.

Zusammenfassung

Die vorliegende Dissertation besteht aus drei empirischen Aufsätzen. Der erste Aufsatz untersucht die Folgen einer Debatte, die einer kollektiven Entscheidung vorausgeht. Dieses Kapitel diskutiert die Ergebnisse eines Lab-in-the-Field-Experiments mit minderjährigen Schüler:innen in Deutschland. Einige von ihnen werden nach dem Zufallsprinzip zugewiesen, im Chat darüber zu diskutieren, wie viel sie an eine Hilfsorganisation spenden möchten, die minderjährige Geflüchtete unterstützt. In unserer Studie führt eine solche Debatte, die einer Abstimmung vorausgeht, zu einer höheren Spendenbereitschaft. Die Debatte hat zwar keinen direkten Einfluss auf das gegenseitige Vertrauen unter den Schüler:innen, die miteinander chatten. Allerdings werden diejenigen, die nach der Diskussion als flüchtlingsfreundlicher wahrgenommen werden, ebenfalls als vertrauenswürdiger von ihren Chatpartner:innen eingestuft.

Der zweite Aufsatz beschäftigt sich mit dem Entstehungsprozess von Überzeugungen über andere Menschen und geht der Frage nach, wie sich ein Verstoßen gegen eine soziale Norm auf die Vorstellungen von Menschen auswirkt, was andere in der gleichen Situation tun würden. Dieses Kapitel basiert auf den Ergebnissen eines Online-Experiments mit einer repräsentativen Stichprobe der britischen Gesamtbevölkerung. Dabei werden die Studienteilnehmer:innen entweder mit einer hohen oder einer niedrigen Versuchung zu lügen konfrontiert. Die Ergebnisse zeigen, dass Proband:innen in dem Umfeld, in dem Lügen attraktiver ist, mehr lügen und als Folge pessimistischere Vorstellungen davon haben, wie sich andere in der gleichen Situation verhalten würden. Die Studie beinhaltet zusätzliche experimentelle Bedingungen, um zu zeigen, dass die beobachteten Effekte nicht durch rationale Erwartungen verursacht werden. Das Experiment liefert kausale Evidenz dafür, dass das Verstoßen gegen eine soziale Norm dazu führt, dass Menschen ihre Vorstellungen über das Verhalten anderer Menschen strategisch verzerren, um ein positives Selbstbild aufrechtzuerhalten.

Der dritte Aufsatz analysiert eine Situation, in der Menschen von einem Algorithmus

unterstützt werden. Die Erkenntnisse basieren auf einem Online-Experiment mit amerikanischen Proband:innen, die wiederholt eine Schätzung abgeben und dabei (meist verzerrte) Ratschläge von einem Algorithmus erhalten. Unsere Studie umfasst zwei Interventionen. Zum einen zeigen unsere Ergebnisse, dass eine abstrakte Erklärung der Funktionsweise des Algorithmus dazu führt, dass die Menschen dem Vorschlag unseres Algorithmus weniger folgen. Gleichzeitig verbessert diese Maßnahme die menschliche Leistung bei der Schätzaufgabe nicht. Ein Offenlegen der richtigen Antwort nach jeder Runde der Schätzaufgabe führt hingegen dazu, dass die Menschen dem Ratschlag unseres Algorithmus weniger vertrauen und die menschliche Leistung bei der Schätzaufgabe sich verbessert. Zum anderen steht unsere Forschung im Zusammenhang mit existierenden Studien, die nahe legen, dass Menschen aufhören, den Ratschlägen eines Algorithmus zu folgen, nachdem sie beobachtet haben, wie der Algorithmus eine falsche Vorhersage generiert. Dieses Ergebnis wird in unserer Studie nicht bestätigt. Die wahrscheinliche Erklärung dafür ist, dass Menschen in unserem Setting nachvollziehen können, aus welchem Gründen einzelne Schätzungen des Algorithmus ungenau sind. Zusammenfassend liefern die drei Aufsätze neue Erkenntnisse über (verhaltens-)ökonomische Konzepte von Vertrauen, Großzügigkeit, der Entstehung von Überzeugungen und der Akzeptanz von Ratschlägen von Algorithmen mithilfe experimenteller Methoden.

Erklärung

Hiermit erkläre ich, Jan Biermann, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Ort/Datum

Unterschrift Doktorand

Eidesstattliche Versicherung

Ich, Jan Biermann, versichere an Eides statt, dass ich die Dissertation mit dem Titel:

„Of Discussions, Beliefs, and Algorithms: Essays in Experimental Economics“

selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach § 6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 18. Januar 2017 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

Ort/Datum

Unterschrift Doktorand

Selbstdeklaration bei kumulativen Dissertationen

Konzeption / Planung: Formulierung des grundlegenden wissenschaftlichen Problems, basierend auf bisher unbeantworteten theoretischen Fragestellungen inklusive der Zusammenfassung der generellen Fragen, die anhand von Analysen oder Experimenten / Untersuchungen beantwortbar sind. Planung der Experimente / Analysen und Formulierung der methodischen Vorgehensweise, inklusive Wahl der Methode und unabhängige methodologische Entwicklung.

Durchführung: Grad der Einbindung in die konkreten Untersuchungen bzw. Analysen.

Manuskripterstellung: Präsentation, Interpretation und Diskussion der erzielten Ergebnisse in Form eines wissenschaftlichen Artikels.

Die Einschätzung des geleisteten Anteils erfolgt mittels Punkteinschätzung von 1 – 100%.

Für mindestens einen der vorliegenden Artikel liegt die Eigenleistung bei 100% .

Für den ersten Artikel (chapter 2) liegt die Eigenleistung für

das Konzept / die Planung bei 25%

die Durchführung bei 35 %

die Manuskripterstellung bei 45 %

Für den zweiten Artikel (chapter 3) liegt die Eigenleistung für

das Konzept / die Planung bei 100%

die Durchführung bei 100 %

die Manuskripterstellung bei 100 %

Für den dritten Artikel (chapter 4) liegt die Eigenleistung für

das Konzept / die Planung bei 45%

die Durchführung bei 45 %

die Manuskripterstellung bei 70 %

Die vorliegende Einschätzung in Prozent über die von mir erbrachte Eigenleistung wurde mit den am Artikel beteiligten Koautor:innen einvernehmlich abgestimmt.

Ort/Datum

Unterschrift Doktorand