Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change

Pace, D.D.

Publication date
2023

Document Version
Final published version

The thesis can be ideally divided into two parts. The first, comprising chapters 1 to 3, consists of basic research on the foundations of human behavior. The second part, which consists of the last chapter, uses state-of-the-art techniques from behavioral and experimental economics to investigate a question relevant to climate policy.

Davide Domenico Pace holds a BSc in Economics of Banks, Insurances, and Financial Intermediaries from the University Milano Bicocca and an MPhil degree in Economics from the Tinbergen Institute. In 2018, he joined the Center for Research in Experimental Economics and Political Decision Making (CREED) at the University of Amsterdam as a PhD student under the supervision of Joël van der Weele and Joep Sonnemans. Davide currently works as an Assistant Professor at LMU Munich.

Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change Davide Domenico Pace

824

tinbergen institute
GRADUATE PROGRAM | Erasmus VU

Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change

Davide Domenico Pace

University of Amsterdam

Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change

ISBN: 97 890 361 0719 8
Cover illustration: Davide Domenico Pace using Midjourney
Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **824** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found here.

Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 6 september 2023, te 16.00 uur

door Davide Domenico Pace
geboren te Milano

Promotiecommissie

Promotores:	prof. dr. J.J. van der Weele	Universiteit van Amsterdam
	prof. dr. J.H. Sonnemans	Universiteit van Amsterdam
Overige leden:	dr. G. Romagnoli	Universiteit van Amsterdam
	prof. dr. T. Buser	Universiteit van Amsterdam
	prof. dr. J.B. Engelmann	Universiteit van Amsterdam
	prof. dr. K. Nyborg	University of Oslo
	prof. dr. F. Zimmermann	Briq

Faculteit Economie en Bedrijfskunde

Acknowledgement of financial support The research for this doctoral thesis received financial assistance from NWO VIDI grant 452-17-004 awarded to Joël van der Weele; the Research Priority Area Behavioral Economics, the Amsterdam Center for Behavioral Change, the 'A Sustainable Future' platform of University of Amsterdam; the Deutsche Forschungsgemeinschaft through CRC TRR 190; the Bavarian Academy of Sciences and Humanities.

Acknowledgments

No good thing in life is done while being alone. This thesis is no exception. Its four chapters only exist thanks to an endless stream of teachings, discussions, conversations, exchanges, and shared experiences that intensively followed each other. Hard to say when it all began. A starting point might be the day Dad read aloud a chapter from “The Paradoxicon" by Nicholas Falletta. I don’t remember how old I was, but surely I could not yet read much alone. The Paradoxicon attracted me mostly for its reproductions of the Escher’s lithography with its infinite stairs and water flows defying gravity. One afternoon Dad pushed beyond those pictures and read me the chapter dedicated to the prisoner dilemma. The Nash equilibrium was completely beyond the understanding of my child’s brain. I simply couldn’t believe that the solution to the problem was for the two prisoners to accuse each other and end up spending five years in jail when they could have both stayed silent and gotten free after one year. Viscerally, I was sure there was something wrong. Little I knew that similar intuitions inspired the work of many behavioral economists.

So, the first huge “Thank you" goes to my family. Mamma, Papà, and Guido, you have always encouraged and supported any initiative, idea, and life project. You infected me with your idealism but also taught me to be practical and to get done what needs to be done. You are always ready to lift the mood with a joke when the times are tough or when I get too serious. Thảo, love of my life, thank you for being always and simultaneously both supportive and critical; for the many adventures we are having around the world; and for all the ideas, stories, and reflections you share with me day after day. And thank you for improving so many research projects with your brilliant comments.

Throughout the years, I had the luck and the privileged to meet many *Maestri* whose teaching still shapes my work and thinking. I am indebted to Paola Biscari for showing me how to study. To Luca Ferraiuolo for explaining again and again that even the best ideas don’t go far if they are presented poorly. To Marco Pernich, from whom I learned how to present and teach. To Antonella Alvino, Ida Sassi, Alfredo Di Legge, and Simonetta Cadirola for all the general knowledge they handed down and without which my research would stand on shaky grounds. To Luca Stanca and Stefania Ottone, who introduced me to behavioral economics. And to my PhD advisors Joep Sonnemans and Joël van der Weele. Joep, I knew your door was always open, and I could always count on you for guidance and advice. Joël thank you so much for entrusting me with your ideas, for leading by example, for always listening, for all the support in the hard times, for giving me space when I had to grow into an independent researcher, and for kindly raising a red flag when you felt I was going in a wrong direction. I am proud of the work we

have done together, sure that we will come up with new creative ideas to push the knowledge frontier, and just happy to be able to call you a friend.

I am also grateful for an amazing group of co-authors, without whom this thesis would not have been written. Thank you to Dianna, who walked the bridge between Neuroscience and Economics and shared all her knowledge on people's decision processes and process tracking techniques. To Taisuke for his wizardry with data and his astounding precision. To Peter for his sharp design ideas that marked fundamental turning points in the life of many projects. Thank you to Andrej and Klaus for a new and exciting collaboration. But most importantly, thank you all for the meetings and discussions that made doing research way more fun.

Over the years, I have been sharing the path with many friends that enlightened the way. Special thanks to Davide, Rudy, and Ilaria for the many hours spent discussing life, love, philosophy, politics, and arts. Thank you for always reminding me that the international, English-speaking, highly educated, hyper-dedicated community I became part of is just one shiny fragment of a complex and polychromatic world. In this shiny fragment, I was lucky to find extraordinary people. A big shout-out goes to Kathi and Andi. I was so fortunate to share with you the TI years, an office during the PhD, and many board games nights. Another huge "Thank you" goes to Chris. If the lockdown months were not so bad, it is mostly because we shared meals after meals discussing science and watching TV series. Thank you to Wally, Eszter, Marco, Adam, Rik, and Aslı for the many coffees, meals under the sun, and parties. And to Ivan, Jeroen, and Stefania for the many kilometers we shared cycling and chatting in the countryside.

Thank you to all the CREEDers for creating a friendly, supportive, productive, and inspiring research group that gives so many opportunities to the PhDs to learn, grow, and have fun in the process. In particular, thank you to Theo, Giorgia, and Jan Engelmann for their sharp feedback and advice. Thanks to Junze, Silvia, Kostas, Jan Hausfeld, Alejandro, Aljaz, Maria, Linh, Johan, Margarita, Nils, Ailko, Ayşe, and Oda for the many chats and foosball games.

I would also like to thank some people at the UvA and TI, who took care of all the admin needed during the PhD. Arianne, thank you for all the support during the job market. Ester and Judith, thank you for your help during my two years at TI. Wilma and Robert, you made sure everything was working smoothly at the UvA and were always ready to help.

Finally, thank you to the Microgroup at LMU where my journey continues and where this thesis took its final shape. Thanks to Florian, Simeon, Matthias, Valeria, Sili, Anik, and all the PhD students for welcoming me so warmly in Munich. I am looking forward to all the retreats, discussions, and new projects we will do together.

Contents

Relationship between the chapters

A common thread unifies the chapters of this thesis in Behavioral Economics: a focus on human beliefs.

The first two chapters ask how people develop their beliefs about what is a fair allocation of resources. The chapter “Fair Shares and Selective Attention" focuses on the role of visual attention in the development of self-serving biases in redistribution. It finds that economic advantage changes how people pay attention to the determinants of success. Furthermore, it shows that shocking people’s attentional patterns changes their preferences for redistribution. These findings suggest that attention-based policy interventions, like schooling and advertising, may be effective in reducing polarized views on inequality.

The chapter “Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms" looks at views of fairness about redistribution as well, but it focuses on the psychological antecedents of self-serving biases. It shows that economically advantaged people shift their beliefs about what constitutes a fair allocation. People also shift their beliefs about the relative importance of luck and merit for economic success, but not their perceptions of what others think is fair. Finally, the paper finds that beliefs about what constitutes a fair allocation play a significant mediating role in developing self-serving biases.

The third chapter focuses instead on people’s beliefs about their memory abilities. It shows that people can be both under and overconfident about how good their memory is depending on the complexity of the memory task. These findings are essential to understand when memory mistakes can explain behavioral biases and when, instead, people might waste resources to remember information they are already likely to recall accurately.

Finally, the last chapter investigates people’s beliefs about the environmental consequences of their actions. It finds that people underestimate the amount of CO_2 associated with common consumer products. The same people that underestimate these emissions care about the climate, leading to the prediction that correcting people’s misperception will cause them to reduce their consumption of the most polluting products. The chapter tests this hypothesis with an experiment on beef consumption, one of the products for which we predict the largest effects of information. The results do not support our predictions. People change their beliefs when told how much beef is polluting, but there is no evidence that they modify their behavior. The results call into question the potential of even carefully-targeted information to affect individual climate action and have implications for the literature on CO2 misperceptions and labeling.

The thesis can also be ideally divided into two parts. The first, comprising chapters 1 to 3,

consists of basic research focused on understanding the foundations of human behavior. The second part, which consists of the last chapter, uses state-of-the-art techniques from behavioral and experimental economics to investigate a question relevant to climate policy. This ideal division inspired the title of the thesis.

Chapter 1

Fair Shares and Selective Attention

This chapter is based on Amasino, Pace and van der Weele (2021)

1.1 Introduction

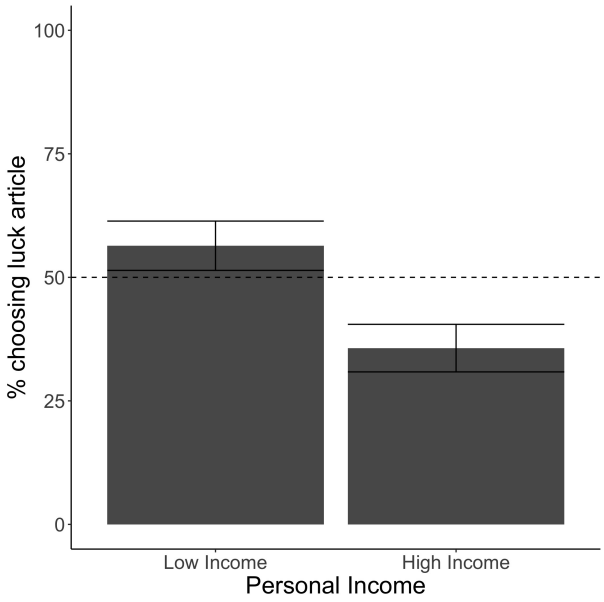
Elites often find ways to justify their economic advantage. Across countries, higher incomes correlate with stronger condemnation of "blue collar crimes" like benefit fraud and weaker condemnation of "white collar crimes" like tax evasion (Ostling, 2009). Affluent Americans are more likely than average Americans to believe that inequalities result from hard-work and intelligence rather than from luck (Suhay, Klasnja and Rivero, 2020), and less likely to redistribute income than the general population (Cohn et al., 2019). The effect of economic privilege is causal: the accidental allocation of land-titles can lead to more pro-market views (Di Tella, Galiant and Schargrodsky, 2007), and the random allocation of an economic advantage to laboratory subjects causes them to redistribute less to unfortunate peers (Konow, 2000; Deffains, Espinosa and Thöni, 2016). In contrast, random shocks that worsen people's economic situations, like sickness and disability, increase the moral appeal of equality (Hvidberg, Kreiner and Stantcheva, 2020). These diverging views about the origin of economic success have been linked to recent political conflict in Western societies Sandel (2020); Gethin, Martínez-Toledano and Piketty (2021).

In this paper, we study the role of visual attention in the formation of attitudes towards merit and redistribution. Attention matters since it is the filter through which people understand their environment, and may depend on an individual's background. For instance, citizens of different socio-economic status may pay attention to news media that provide different narratives about the nature and origin of inequality. We ask how socio-economic status shapes attention to the role of merit and luck, and how such attention affects concerns for fairness and redistribution. The answers to these questions can provide policy levers to combat bias and polarization in attitudes towards meritocracy and economic success and help understand the competition for attention by activists and politicians.

Before describing our main investigation, we motivate our research questions with survey evidence on the relation between socio-economic status and attention. In an online survey

($N = 767$), we asked respondents from different income groups to read one of two articles titled "Luck looms larger in success than most of us think" and "Why high earners work longer hours". We expected that people with high socio-economic status would be more reluctant to learn about the role of luck, and hence less likely to attend to the "luck-article", as it may raise doubts about the merits of their relatively higher income. Indeed, Figure 1.1 shows that only 35.7% of high income participants chose to look at the luck article, compared to 56.4% of the low income participants (χ^2=32.09, $p < 0.001$). Higher income also has a strong, negative correlation with positive attitudes towards redistribution (Kendall rank correlation $\tau = -0.306$, $p < 0.001$).[1]

Figure 1.1: Choice to learn about the role of luck by income level.



Choice of article split by income level, with Low Income defined as $<$ £10.000, and High Income as $>$ £70.000. The Y-axis shows the percentage of participants choosing the article titled "Luck looms larger in success than most of us think" instead of the one titled: "Why high earners work longer hours". The error bars represent 95% confidence intervals.

These results suggest an interplay between economic status, attention to merit and luck, and attitudes towards redistribution. We rigorously investigate the causal links between these variables in our main study, consisting of a series of large online experiments ($N = 1500$). In a design inspired by Konow (2000), participants first produce a surplus by providing correct responses in a series of real effort tasks. In two "Status" treatments, we create "Advantaged" and "Disadvantaged" subjects by explicitly randomizing half of the subjects to a higher pay rate per correct response. Subsequently, a subset of the subjects assume the role of "dictator" ($N = 600$) and divide the surplus generated by two participants, one with Advantaged status and one with Disadvantaged status, in a sequence of allocation tasks. In the "Involved" condition, the dictator is one of the participants who generated the surplus. In the subsequent "Impartial" trials, the dictator divides the surplus generated by two other participants.

[1]Details about the implementation and outcomes of the survey are in Online Appendix A.1.1.

Before dictators make their allocations, we measure their visual attention to the sources of the surplus. Dictators can uncover two sources of information. First, "outcome" information mirrors information most typically available to us. It shows the total contribution of each participant to the surplus, thus combining merit (correct answers) and luck (the randomly determined pay rate). Second, "merit" information shows the number of correct answers of both participants, thus providing a measure of performance net of the aleatory pay rate. We measure the visual attention to these two sources with the tool MouselabWEB, tracking how each subject moves their mouse over the screen to uncover different types of information (Willemsen and Johnson, 2019).

We focus on visual attention or "dwell time" because it is the key locus of competition for attention and because salient contextual elements that affect gaze patterns have been shown to affect choice (Krajbich, 2019; Orquin and Mueller Loose, 2013; Bordalo, Gennaioli and Shleifer, 2021). To understand the (causal) role of dwell time in dictator's decisions, we implement three "Focus" treatments. In the "Free Focus" treatment, participants face no restrictions on their attention. In contrast, the "Merit Focus" and "Outcome Focus" treatments impose restrictions on the time that can be spent looking at different types of information, enabling participants to pay more attention to the merit or outcome. This manipulation changes behavior because it affects the relative time spent on merit and outcome information and not because it makes easier for participants to be willfully ignorant as in Dana, Weber and Kuang (2007). We further address the difference between our study and the willful ignorance literature in the discussion section.

The results show strong evidence of self-serving bias: compared to Disadvantaged dictators, Advantaged ones keep a larger share of the pie in the Involved condition. They also allocate more to other Advantaged recipients in the Impartial trials where dictator's own income is not at stake, replicating results from Konow (2000). This result indicates that the experience of economic advantage changes allocation behavior beyond narrow self-interest.

We then turn to our main interest: the role of attention. First, we find evidence for selective attention: compared to Disadvantaged dictators, Advantaged ones pay relatively more attention to outcome information, which incorporates the random differences in pay rate that favor the Advantaged participants. By contrast, Disadvantaged dictators pay more attention to merit information, which is based on performance only. This pattern arises over multiple trials in the Involved decisions and persists in subsequent Impartial decisions.

Second, and perhaps most importantly, we find that attention plays a causal role in redistribution decisions. The Outcome Focus treatment, which encourages people to look longer at contributions that include the luck component, increases the share of the pie going to Advantaged recipients compared to the Merit Focus treatment. This effect of attention is particularly pronounced among Advantaged dictators. The effect of attention is substantial: making dictators look one second longer at merit versus outcome information (that is, redirecting, about a quarter of average dwell time), reduces the impact of having an advantaged position on allocations by 40% when dictators own income is at stake. We show that this effect is driven by changes in dwell time, and not by completely avoiding some information as in previous literature on the topic (e.g. Dana, Weber and Kuang (2007)). We can also rule out experimenter demand effects

or processing errors as psychological mechanisms behind the results. Instead, we show that attention causes subjects to change their views of what is appropriate or fair in these division problems in ways that carry over to the Impartial trials.

Relative to previous literature on redistributive attitudes, which we survey in more detail below, our focus on attention allows us to study the cognitive underpinnings of self-serving bias. We show that attention plays a causal role in redistribution and fairness decisions, and attention-based interventions are effective as a lever to influence such decisions. This opens a new window on socio-economic cleavages in attitudes towards meritocracy and redistribution, and provides a starting point for interventions to reduce bias, not just in redistributive decisions, but also in other domains where discrimination of disadvantaged groups plays a role.

1.2 Literature Review

Our research relates to a several strands of literature. First, we contribute to a behavioral literature on the role of merit in redistribution. A number of laboratory experiments shows that participants are more willing to redress inequalities based on luck rather than merit (Krawczyk, 2010; Cappelen et al., 2013; Durante, Putterman and Van der Weele, 2014; Lefgren, Sims and Stoddard, 2016; Cappelen et al., 2017; Bortolotti et al., 2017; Buser et al., 2020). Almås, Cappelen and Tungodden (2020) have shown that this tendency is robust across countries, even if there are differences in the overall tendency to redistribute, although Jakiela (2015) finds that the distinction between merit and luck less strong in rural villages with strong egalitarian norms. Piff et al. (2020) show that priming people with situational rather than dispositional attributions for poverty causes an increase in egalitarianism. We add to these insights by showing that attention to merit and luck is endogenous and has a causal effect on the allocation of an economic surplus.

Second, we contribute to an understanding of well-documented self-serving biases in redistribution. In particular, the seminal paper by Konow (2000) identifies a self-serving bias exhibited by players with a randomly-assigned advantage who give more to themselves and also to other advantaged players, even when their own income is not at stake. Rodriguez-Lara and Moreno-Garrido (2012) and Deffains, Espinosa and Thöni (2016) use similar designs and replicate these main results. Espinosa, Deffains and Thöni (2020) show that the bias is robust to ex-post information provision, highlighting the role of luck in the formation of inequality. Several papers, cited in the introductory paragraph, demonstrate self-serving bias outside the laboratory; other forms of self-serving bias have been found in a wide range of domains (Bénabou and Tirole, 2016). While this literature demonstrates the importance and self-serving nature of fairness views, it has treated the formation of such beliefs largely as a black box. Our paper opens the box by focusing on the role of attention, opening new channels for policy interventions.

Third, our focus on attention contributes to a fast-growing literature on the role of attention in economic decisions. In particular, we relate to a literature that links choice to various attentional mechanisms (surveyed in Engelmann, Hirmas and van der Weele, 2021; Fisher,

2021). First, goals and preferences can direct "top-down" attention to the more highly-valued options during choice. We expand this literature to look at redistributive decisions, showing how economically advantaged decision makers look at information that is more "convenient". This freely-directed attention to more appealing information is in line with top-down attention. Second, attention can also be captured in a "bottom-up" manner, where the "salience" of contextual elements affects attention and decisions. This approach has been modeled to explain various deviations of economic rationality (Shimojo et al., 2003; Bordalo, Gennaioli and Shleifer, 2012, 2021). In the choice literature, top-down attention is understood to drive a large part of choice, but bottom-up salience and random fluctuations in attention have also been found to matter, especially for more difficult choices where the options are closer in value (Milosavljevic et al., 2012; Smith and Krajbich, 2018). For these more difficult choices, relative dwell time on options or attributes can impact choice (Krajbich et al., 2012; Konovalov and Krajbich, 2016; Fisher, 2021; Pärnamets et al., 2015; Mullett and Stewart, 2016; Smith and Krajbich, 2019). In our study, we manipulate attention in a way that still preserves top-down attention, but also may act on bottom-up attention by making certain information relatively easier to access. Such attentional manipulations may act primarily on "difficult" or conflicted decisions. Finally, a newer area of research suggests that attentional history or habits can drive future attention, but its role in complex choice tasks has just started to be explored (Theeuwes, 2019; Jiang and Sisk, 2019; Gwinn, Leber and Krajbich, 2019). We examine how attentional patterns developed in choices with one's own payoff at stake spillover into impartial decisions, which relates to this literature on attentional history and habits. Our study contributes to this literature, by showing that manipulating dwell time affects monetary allocations in self-other and other-other decisions.

Finally, we relate to an emerging literature on the role of attention in pro-social decisions. One such line of research has focused on the phenomenon of information avoidance and selective search in moral situations (Dana, Weber and Kuang, 2007; Grossman and Van der Weele, 2017; Chen and Heese, 2021). In these studies, participants choose whether or not to reveal information about the consequences of their decisions on others. A substantial number of participants avoid such information, maintaining their self-perception as a moral person while making a selfish choice, something that they would be unable to do if confronted with the consequences of their choice. In such a setup, it is impossible to act on the information that is avoided, leaving little nuance for understanding how people sort through the barrage of conflicting information experience outside of the lab.

In contrast to this binary reveal/avoid decision, we look at a more continuous measure of attention, namely dwell time. This setting is more realistic in capturing situations where people are exposed to many different perspectives and types of information. Indeed, our study shows that avoidance is very low, and that even if people reveal all information about payoffs, the *length* of the relative dwell time on that information affects their choice as it affects the weight on different types of information.

A separate line of attention literature in social decision-making uses eye-tracking and mouselab technology to focus on continuous attention, but without the clear distinctions of merit and luck in determining fairness. Fiedler et al. (2013) show correlations between eye movements and

social preferences in social allocation problems. These correlations are replicated in mouselab-WEB by Bieleke, Dohmen and Gollwitzer (2020). Further, participants adjust their gaze to appear prosocial or take others payoffs more into account in strategic settings where their payoffs depend on others' decisions (Fischbacher, Hausfeld and Renerte, 2020). Ghaffari and Fiedler (2018) look at the causal, bottom-up effect of attention. Replicating and extending Pärnamets et al. (2015), they manipulate attention to payoffs in a social allocation problem by interrupting the decision-making process after subjects look at a certain option for a pre-determined amount of time. This exogenous variation can explain about 11% of the variation in visual attention and about 1% of changes in choice. Other results have shown correlations of attention with loss-framing (Fiedler and Hillenbrand, 2020) and in-group bias (Rahal, Fiedler and De Dreu, 2020; Fischbacher, Grammling and Hausfeld, 2021) in social dilemmas.

Our approach differs from the empirical studies cited above, and all attention-tracing studies in this domain that we are aware of. Instead of measuring attention to the payoffs in an economic game, we study attention to the *determinants* of economic production and show how this affects distributive decisions. Thus, it is one of the first papers to link attentional processes with the reasoning behind fairness judgments, elucidating the origins of (self-serving) fairness views. The most closely related paper to this endeavor is Waldfogel et al. (2021), one of the few studies on attention towards economic inequality. They show that political ideology affects whether people detect inequalities in everyday situations, whereas we focus on the determinants of inequality.

1.3 Design

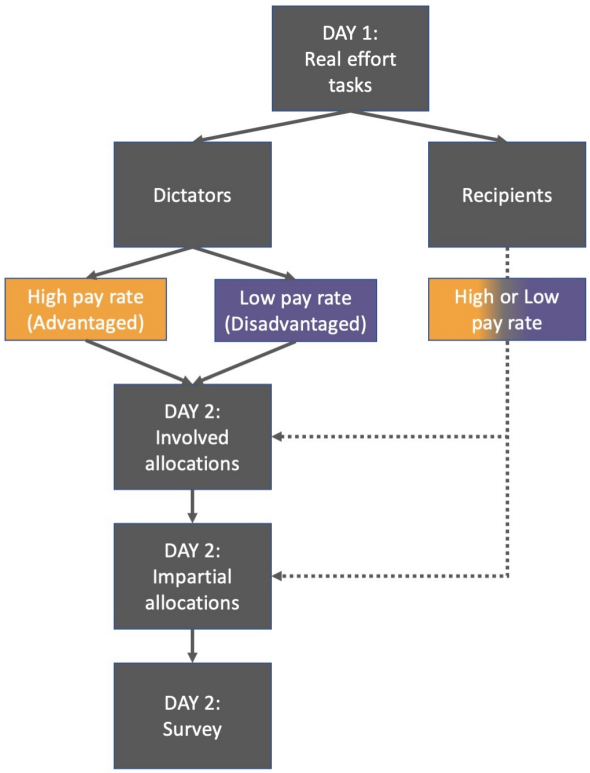
The study consists of two orthogonal treatment dimensions, leading to a 3×2 design, with 100 decision-makers (dictators) in each cell, as outlined in Table 1.1. The data were gathered in two experiments. Experiment 1 generated the data for the Free Focus treatment. It aims to a) replicate previous findings on the relationship between economic status and attitudes toward redistribution and b) establish a causal relationship between economic status and attention. Experiment 2 generated data for the Merit and Outcome Focus treatments, and allows us to c) investigate the causal relationship between attention and attitudes towards redistribution.

Table 1.1: Overview of treatments and number of dictators

Status	**Attention**		
	Free Focus	Merit Focus	Outcome Focus
Advantaged	100	100	100
Disadvantaged	100	100	100

Overview of the treatments in our 3×2 design. The data for the Free Focus treatments comes from Experiment 1. The data from the Merit and Outcome Focus treatments come from Experiment 2. The numbers in the cells indicate the number of dictators per treatment.

Figure 1.2: Timeline for Day 1 and 2 for Both Experiments.



Each experiment happened over 2 days: on Day 1, participants completed real effort tasks to generate a surplus, and on Day 2, participants in the role of dictators divided the surplus. Figure 1.2 displays the timeline shared by the two experiments.

For Experiment 1, we recruited 200 dictators and 300 recipients from Prolific.co. The data was collected between the 13th and 19th of July, 2020. For Experiment 2, we recruited 400 dictators and 600 recipients from Prolific.co.[2] The data was collected between the 23rd and 30th of November, 2020. Across both experiments, we paid a completion fee of $£2.85$ for Day 1 and $£6.15$ for Day 2 plus an average bonus of around £3 per participant. Overall, 1500 participants completed the study in the role of either dictator or recipient. In our analyses, we focus on the attention and allocation decisions of the 600 participants assigned the role of dictator.[3]

56% of the participants are man and the average age is 24 years old. Online Appendix A.1.2 gives further details about the subjects' demographic characteristics and about the recruitment procedure. Furthermore, it shows that attrition between the two experimental days is minimal and balanced across Status treatments.[4].

[2]We recruited more recipients than dictators because in the Impartial trials the dictators split the amount generated by two recipients.

[3]None of these dictators took then part in the motivating survey discussed in the introduction.

[4] Attrition by dictators from Day 1 to Day 2 was low because participants had to complete both days to be paid. In total, 11 participants from Experiment 1 and 15 from Experiment 2 had to be replaced. Of these 26 dictators, 10 did not start Day 2 and, hence, dropped out before knowing their Status. The other 16 started Day 2, they learnt about their Status, but they did not complete the experiment. 6 of these 16 dictators were Advantaged and 10 Disadvantaged. The difference in attrition rate between the two groups is not significant (Fisher's exact test $p = 0.45$).

1.3.1 Day 1: Surplus Generation

On Day 1, participants completed 8 sets of real effort tasks. In each task set, participants had a limited time period to complete as many tasks as possible. There were 4 different types of tasks: moving sliders to a predetermined position, logic questions, counting the number of zeros in a table, and solving Raven's matrices. The 8 task sets were evenly split among the different task types. In every task set, each correct answer earned a monetary reward. When completing the task sets, participants did not know the exact monetary reward they would receive. However, they knew that they would randomly be assigned a high or low pay rate per correct answer, the amount of both pay rates, and that they would learn which pay-rate applied to them at a later stage. The high pay rate was always 3 times the low pay rate, but pay rates were calibrated (based on pilot data [5]) according to task type to result in an average surplus of $\pounds 3.5$ per task set.

Similarly, participants were aware that the assignment to a high or low pay rate would apply to all of their tasks. We checked participants understanding of the randomness and persistence of the pay-rates with two comprehension questions, which they had to get correct to continue with the study. Participants were also informed that they would be paired with other participants and their earnings would go into a single common account but did not know how this would be divided.

We informed participants about the two possible pay-rates and about the existence of the common account to provide incentives for exerting effort and, at the same time, be transparent at all stages of the study. Transparency is especially important towards the recipients as they would not continue to Day 2. Since all participants were given the same information and were not informed of their pay rate at this stage, the information should not affect participants differentially.

1.3.2 Day 2: Surplus Division

After the Day 1 surplus generation was complete, we split participants into dictator and recipient roles. Only the dictators were invited to Day 2, which started one day after Day 1. Day 2 was divided into 3 parts. In part 1, dictators split earnings between themselves and recipients, termed "Involved" allocations. In part 2, they split earnings between pairs of recipients, termed "Impartial" allocations. In part 3, they answered questions about their strategies, beliefs, and perceptions of norms.

At the beginning of Day 2, dictators learned their pay-rate per correct answer. We call participants who received the high pay rate "Advantaged" and those with the low pay rate "Disadvantaged," and we refer to this difference as the "Status" treatment. Participants then received instructions for the Involved allocation task. The joint earnings of a pair in a task were merged into a common account, and the dictator chose how to allocate this common account between themselves and the paired recipient. Over 20 trials, the dictators were matched with different recipients, with one of the 8 task sets underlying the common account in each of the trials.

[5] The pilot included 50 dictators with only allocation behavior (no attention data) and was collected February, 2020.

We matched Advantaged dictators with disadvantaged recipients and vice versa Disadvantaged dictators with advantaged recipients. The dictators were made aware of these inequalities in the instructions, and we checked their understanding with a comprehension question. The explicit and consistent allocation of relative advantage throughout the experiment mimics systematic advantages like those due to the socioeconomic position of parents. It allow us to investigate how such advantages affect attention to merit and luck information. We created trials such that dictators outperformed recipients on 50% of the trials to make sure that the effects of pay rate and relative performance were not confounded. During each trial, dictators received information about how the common account was generated (detailed in the next section) and made their allocation decisions.

In the next part of Day 2, dictators made Impartial allocation decisions for two recipients. Just as in the Involved allocations, the Impartial allocations always included one Advantaged and one Disadvantaged recipient. Over 20 trials, dictators chose how to divide the common account produced by pairs of different recipients. Participants always completed the Involved trials before the Impartial trials in order to test whether self-serving biases developed in Involved decisions persisted into Impartial decisions, as in Konow (2000). This order was chosen deliberately: putting the Involved trials first gives subjects experience of their economic status. This mirrors situations outside the laboratory where people have a lifetime of experience in their economic roles. The status in the Involved condition thus functions as an experimental treatment to investigate the bleed-over of fairness rules and attentional habits into impartial decisions.[6]

Decisions were incentivized by implementing one of each dictator's 40 decisions. The average surplus per pair of participants in each task was £6.99 in Experiment 1 and £7.10 in Experiment 2. These amounts are approximately 1.4 times the minimum hourly wage on Prolific at the time of the study, so the allocation decisions had reasonably high stakes. If the decision came from the Involved allocations, the dictator received a bonus payment equal to the amount they kept for themselves, and the recipient received the amount allocated to them. If the decision came from the Impartial allocations, the dictator received $\pounds 1$ and each of the two recipients received what the dictator allocated them.[7]

[6]The fixed order that allows us to examine spillover from Involved to Impartial decisions also limits the interpretation of Impartial allocations because there is a time-confound between later decisions and decision-type. The results might be different if Impartial decisions were made first. For example, if participants weight their own status less in Impartial decisions and there is cognitive dissonance to shifting fairness strategies, this could reduce the self-serving bias also in the Involved decisions, leading to overall more similar attention and allocations regardless of Status. Alternatively, participants could shift their fairness rules even in Impartial decisions if they anticipated the effect on Involved decisions. Such order effects have been investigated in allocation decisions without luck by Dengler-Roscher et al. (2018) with some evidence suggesting that putting Impartial decisions before Involved reduce self-serving bias.

[7]We pre-assigned which type of trial (involved or impartial) would be relevant for payment, and which recipients would get the bonus to ensure that all dictators and recipients were paid a bonus based on a single allocation decision. Recipients could appear in multiple different dictators' allocation decisions.

1.3.3 Attention Measurement

Before every decision, the dictators could look at information about the way the money in the common account was generated, as illustrated in Figure 1.3. First, dictators could see the amount of money in the common account and the type of task that produced it. All 8 task sets were used approximately equally across the 40 trials. Dictators could spend as much time as they wanted on this screen. Next, dictators had 6 seconds during which they could reveal information about the number of correct questions each participant answered in the task - merit information - and the monetary contribution of each member of the pair to the account - outcome information. Merit and outcome information were chosen as they correspond directly to meritocratic and libertarian fairness criteria, respectively, which are relevant for dictator decision-making (Cappelen et al., 2007; Rodriguez-Lara and Moreno-Garrido, 2012).[8] This information was divided in four boxes labelled with participant and information type. All boxes were initially closed, but participants could open a box by hovering over it with their mouse cursor. Only one box could be opened at any time: when the cursor moved away, the box closed again. This was implemented with MouselabWEB which also allowed us to easily record the number of times each box was open and the amount of time the dictators spent on each box (Willemsen and Johnson, 2019). When the time limit was reached, the page automatically updated to the allocation screen where participants decided how to split the money using a slider.[9]

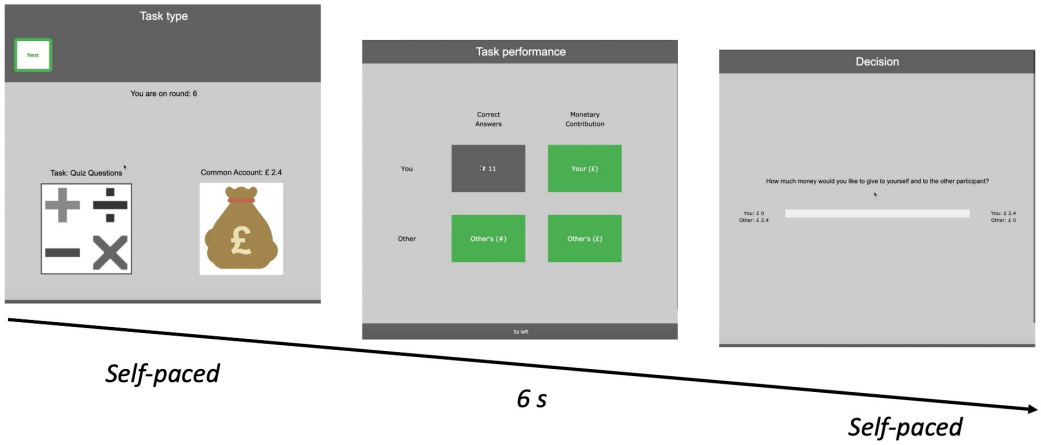
1.3.4 Focus Treatments

We implemented three “Focus” treatments that varied the time different types of information could be accessed. In the “Free Focus" treatment, there was no limit on the number of times a box could be reopened or for how long it could be opened within the overall 6 s time limit. The 6 s time limit was chosen to control for the overall information-gathering period across participants so that differences in attention would be meaningfully comparable. Furthermore, the limit pushes participants to prioritize gathering information that they find relevant and meaningful which also reduces the obligation to reveal or explore all information. Finally, the time limit introduces in the experiment the tight attentional constraints that permeate real life (Gabaix, 2018). The 6 seconds limit is in line with prior research that uses limits as low as 3 s for decisions with 2 pieces of information to understand the impact of attention on choice, doubled to 6 s for 4 pieces of information (Ghaffari and Fiedler, 2018).

[8] In particular, outcome information is reflective of information typically available outside of the lab as it incorporates both merit and luck (for example one’s income can be inferred with some approximation from his/her lifestyle). Merit information isolates the role of merit and separates it from luck. This information is typically not easily available in real life, but can sometimes be obtained with some effort. We exclude pure luck information because the pay differential for Advantaged and Disadvantaged is constant across trials and known in advance.

[9] Given this set-up, one might be concerned that participants don’t need to open all the boxes to obtain the information they need. For example, a participant that remembers the pay differential can calculate merit from outcome and vice versa. However, this is a complex and effortful calculation. In providing all the information, we make it easier for the participants to implement the different fairness rules without relying on their memory and arithmetic abilities. Indeed Section 1.5.5 shows that almost all participants open every box. In any case, if participants indeed calculate the content of the boxes they do not see, we will *underestimate* the effect of attention on allocations.

Figure 1.3: Information sequence



The image shows the sequence of information during allocation decisions. First, participants saw the amount in the common account and the task type that generated the surplus. Next, they had 6 seconds to reveal merit and outcome information by hovering over the boxes with their cursor: The closed green boxes indicate the type of information, and opened boxes are grey with the values inside. Finally, participants made allocation decisions.

The Constrained Focus treatments limited the time participants could see particular information, building on prior work manipulating attention (Pachur et al., 2018; Pärnamets et al., 2015; Ghaffari and Fiedler, 2018). These restrictions were designed to shift dwell times on the different types of information, without making any information unavailable and preventing implementation of any particular decision criterion. In the discussion section, we show evidence that this strategy was successful.

In every trial, two of the four boxes could be opened for no more than 400 ms each. The other two boxes could be opened for no more than 1600 ms each. The total maximum of 4 s spent on box information was chosen to closely match the average time spent on information from the Free Focus experiment, which was 3.8 s. The 400 ms constraint was chosen because information can still be processed and remembered for later use at this timing, whereas timings of 200 ms or lower may be actually restrictive for recognition (DiCarlo, Zoccolan and Rust, 2012; Potter, 1976). Prior attention manipulations have used minimum dwell times of 250 ms and 300 ms (Armel, Beaumel and Rangel, 2008; Pärnamets et al., 2015; Pachur et al., 2018; Fisher, 2021).

Participants are not required to look at any information: they can choose the sequence and which information to reveal, some information is simply available for a longer time if participants choose to reveal it for longer. Boxes could still be opened multiple times within the 6 s time limit, each time counting against the individual box time limit. Participants with these constraints were informed that some boxes might close permanently before the 6 s was over, but they were not informed which boxes would close.

Demand effects and trial-by-trial restrictions. Experimenter demand effects may arise when certain information is made more salient or more readily available, as participants may infer that this information is more "important". To obfuscate the nature of the restrictions and counter such effects, we implemented our main treatment in 14 of the 20 trials in each condition. In the remaining six trials, restrictions were placed on orthogonal box dimensions.[10] Across Involved and Impartial trials and Focus treatments, the order in which the trials with different restrictions appeared were randomized at the individual level.

Our obfuscation strategy was successful, as only a small minority of subjects could identify the box restrictions they faced during the experiment (see Section 1.5.3). In addition, the contrast between the within-subject trial-by-trial changes in restrictions and the sustained between-subject treatment changes allows us to better understand the mechanisms of the attention manipulation (see Section 1.5.4).

1.3.5 Surveys

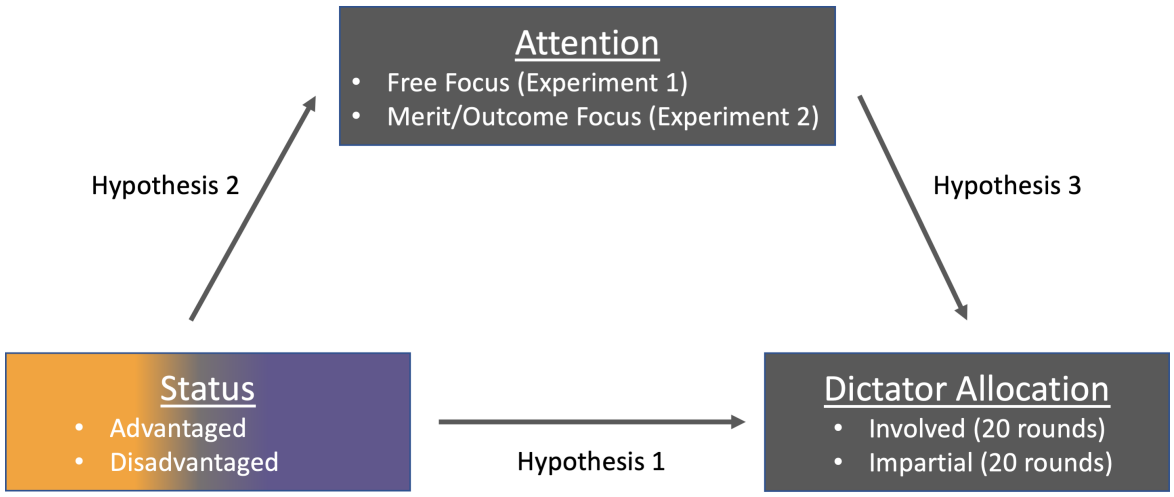
After both experiments, we asked dictators a series of questions about their strategy, their perceptions of various fairness criteria, and their demographics. We asked participants an open-ended question about how they chose to make their allocations. We also asked them to rate the moral appropriateness of dividing according to egalitarian (equal split), meritocratic (effort-based), and libertarian (maintain differences due to effort and luck) criteria, as well as the social norms related to these criteria using the method in Krupka and Weber (2013). Next, we asked them how they thought others would rate these different criteria, overall, and depending on the other's Dis(Advantaged) status. Participants could earn a bonus of $\pounds 1$ for correctly predicting others' answers. We also asked for gender, country, political leaning, education, and income level. In Experiment 2, we additionally elicited incentivized beliefs about some aspects of other participants' performance using the same Krupka and Weber (2013) method and $\pounds 1$ bonus for correct prediction as for social norms.

1.3.6 Hypotheses

Our overall aim is to characterize the role of attention in redistributive decisions and self-serving bias, induced by our Status treatment. To do so, we identify three causal relationships, depicted in Figure 1.4, which drive our research questions and hypotheses. We preregistered these hypotheses on Aspredicted.org in two separate files, one for each experiment, which are included in Appendix A.1.

[10]For instance, in the "Merit Focus" treatment, 14 of the 20 decisions restricted outcome information to 400 ms and merit information to 1600 ms. This enabled participants to look longer at merit. In the remaining six trials, the 400 ms restrictions were placed either on merit information (2 decisions), Advantaged member information (2 decisions), or Disadvantaged member information (2 decisions). In contrast, in the "Outcome Focus" treatment, 14 trials restricted merit information to 400 ms, while the remaining 6 trials split the 400 ms restrictions evenly between the other information dimensions.

Figure 1.4: Framework for the Experimental Design and Hypotheses.



The first relationship involves status and behavior. To understand whether self-serving biases affect fairness decisions, we try to replicate the effects documented by Konow (2000) and follow-up studies (Rodriguez-Lara and Moreno-Garrido, 2012).

Hypothesis 1 (Self-serving bias)**.** *In the Involved condition, Advantaged dictators give less money to the recipients, and more money to themselves, than Disadvantaged dictators.*

The second relationship concerns the impact of status on attention. Following a literature on motivated reasoning (e.g. Kunda, 1990; Bénabou and Tirole, 2016), we expect that dictators in the Involved conditions need a justification for transferring a larger amount to themselves. Selective attention is employed in the search for such justifications. Independently of their performance in the tasks, Advantaged dictators benefit more from looking at and dividing according to outcome information that incorporates their random advantage in pay-rate. In contrast, Disadvantaged dictators may find more justifications in ignoring the luck component in outcome and focusing on merit information, which is purely effort-based. This leads to the following hypothesis.

Hypothesis 2 (Selective attention)**.** *In the Involved condition, Advantaged dictators spend relatively less time on correct answer information and more time on monetary contribution information than Disadvantaged dictators.*

The third and main hypothesis relates to the causal role of attention on behavior, which we address using our attention manipulations in Experiment 2. We expect that increasing the dwell time on merit relative to outcome will lead to a reduction in giving to Advantaged participants. This hypothesis depends on a large body of literature, reviewed in Section 1.2 showing that merit is in an important criterion in redistribution, and that exogenous changes in salience or dwell time can affect choice.

Hypothesis 3 (Attention impacts allocations)**.** *In the Involved condition, increased attention to merit in the Merit Focus condition leads to a reduction in giving to Advantaged recipients compared to the Outcome Focus manipulation.*

Finally, we investigate how much the effects persist in Impartial allocations, where dictators decide between two recipients, and hence their self-interest is not at stake. This is a measure of how much subjects internalized the fairness criteria or attentional habits they formed during the Involved stage.

Hypothesis 4 (Persistence)**.** *The patterns in Hypothesis 1, 2 and 3 continue to hold in the Impartial trials.*

To give further backing to these hypotheses, Section 1.5.1 discusses a formal model of attention and fairness. Following our preregistration, we test all our hypotheses with rank-sum tests, based on the average of individual decisions over all rounds, thus eliminating concerns of dependence of observations. In addition, we use linear regressions controlling for subject characteristics, clustering standard errors by individual.

Attention measures. We measure attention as the dwell time on the two different types of information: merit and outcome information. Dwell time is the focus of most of the literature on visual attention. In section 1.5.5, we look at alternative measures like information avoidance. As a measure of selective attention, we use the *difference* between these two dwell times, which we will shorthand with "ΔAttention", i.e.

$$\Delta\text{Attention} := \text{Dwell time on merit information} - \text{Dwell time on outcome information},$$

where each variable is measured in seconds. To calculate the dwell time on merit (outcome) information, we simply sum up the dwell time on the merit (outcome) for both contributors to the surplus, as the comparison is necessary to make an informed comparison.

In keeping with the literature, in our main analysis, we disregard dwell times when a box is opened for less than 200 ms, as this is considered too short to fully process information (Willemsen and Johnson, 2019; Pachur et al., 2018; DiCarlo, Zoccolan and Rust, 2012). Nevertheless, in Online Appendix A.1.9 we show that our results are robust to using a threshold of 100 ms or including all dwell times regardless of length. Furthermore, in our main specifications, we will not control for the total dwell time of individuals, which is an endogenous regressor that could bias the estimated effect sizes. In any case, in Online Appendix Table A.1.5 we show that our main regression results are robust to the inclusion of this control. All our statistical tests are two-sided, even though our preregistered hypotheses are directional and therefore would have justified a one-sided test.

Table 1.2: Summary Statistics

Panel A: Involved Trials							
		Free Focus		Merit Focus		Outcome Focus	
Allocation		Adv.	Dis.	Adv.	Dis.	Adv.	Dis.
	% given to Adv.	61.5%	50.4%	59.1%	48.3%	64.1%	48.4%
	% given to self	61.5%	49.6%	59.1%	51.7%	64.1%	51.6%
Attention							
	Merit Info (s)	1.68	1.81	1.41	1.33	0.86	0.88
	Outcome Info (s)	2.12	1.90	0.91	0.85	1.56	1.40
	Δ Attention (s)	-0.44	-0.093	0.50	0.48	-0.70	-0.52
Observations		1995	1993	1986	1986	1986	1984
Panel B: Impartial Trials							
		Free Focus		Merit Focus		Outcome Focus	
Allocation		Adv.	Dis.	Adv.	Dis.	Adv.	Dis.
	% given to Adv.	56.3%	52.0%	54.4%	52.5%	56.5%	52.1%
Attention							
	Merit Info (s)	1.96	2.08	1.52	1.36	0.82	0.90
	Outcome Info (s)	1.90	1.57	0.79	0.68	1.30	1.12
	Δ Attention (s)	0.07	0.51	0.73	0.69	-0.48	-0.22
Observations		1994	1990	1987	1988	1978	1986

1.4 Results

We first give an overview of our main treatment effects, before we delve into more details of the different experiments and the interactions between our treatments.

1.4.1 Summary Statistics

We start by evaluating the comparability of the experiments and the engagement of the participants with the merit and outcome information. In the Session 1 production phase, participants exhibited similar performance across Experiment 1 and Experiment 2. On average, participants achieved 13 correct answers per task set in Experiment 1 and 13.5 in Experiment 2, suggesting that participants put effort in completing the tasks in both experiments.

Table 1.2 summarizes the means of the most important outcome variables.[11] First, the share of the surplus given to Advantaged members averaged over both dictator types was 56% for Involved allocations and 54% for Impartial allocations in Experiment 1 and 55% for Involved allocations and 54% for Impartial allocations in Experiment 2. Dictators kept the entire surplus

[11]Each treatment should have 2000 observations, but fewer than 1.5% of observations were not recorded, leading to the varying number of observations. Because the study was conducted online, it is not clear whether these observations were dropped due to an issue with our online database or with participants' computers. However given the number of non-recordings is low and spread across treatments and participants, it is unlikely to affect our results.

for themselves in only 2.8% of the decisions, in accordance with previous findings that dictators respect earned income (Cappelen, Sørensen and Tungodden, 2010; Rodriguez-Lara and Moreno-Garrido, 2012).

Second, participants engaged with merit and outcome information before making their allocations. In the Free Focus treatment, they spent on average 3.8 seconds of the available 6 seconds revealing information in both Involved and Impartial decisions, which amounts to about 2.5 minutes of search time over the entire experiment. Furthermore, pooling across Involved and Impartial decisions, information-seeking was equally distributed between information about correct answers (merit) and monetary contribution (outcome).[12] In Experiment 2, where certain types of information were restricted, participants spent on average 2.3 seconds revealing information in the Involved decisions and 2.1 seconds in the Impartial decisions, also approximately evenly distributed among merit and outcome information pooling across decision types. This is a relatively large reduction in the time spent revealing information compared to endogenous attention in Experiment 1, likely due to the time limits, but participants still engaged with the information nevertheless.

1.4.2 Main treatment effects

Our main treatment effects are captured in Table 1.3, providing a test of our three hypotheses using regression analyses with standard errors clustered at the individual level.

Hypothesis 1: Self-serving bias. Status has a large effect on allocations in the Involved trials. Table 1.3, Column (1) regresses the share allocated to Advantaged subjects on a dummy for the Advantaged. It shows that the Advantaged subjects receive 10 percentage points (roughly 20 percent) more of the pie from the Advantaged dictators (that is from themselves) than from the Disadvantaged dictators (rank-sum test of average allocations $p < 0.001$). The impact of being Advantaged is also apparent in the share dictators kept for *themselves*. For instance, in the Free focus treatment – arguably the cleanest test of Hypothesis 1 – Advantaged dictators kept 61.5% of the pie compared to slightly less than 50% by Disadvantaged dictators ($p < 0.001$, rank-sum test). In fact, the two ways of looking at the division are almost equivalent, because as Table 1.2 shows, the Disadvantaged dictators are very close to splitting the surplus 50-50.

The average division around 50% by the Disadvantaged does not mean they are always splitting the surplus evenly. Allocations by both the Advantaged and Disadvantaged change with the number of correct answers given by each member of the pair. We account for this variable in

[12] We collapse across self and other boxes to focus only on merit and outcome information because these are our variables of interest as described in our hypotheses. Furthermore, there is evidence that participants look at information in an attribute-wise manner, comparing merit for self and other or outcome for self and other. The Payne Index (the proportion of option-wise (within self-performance or within other performance) transitions minus attribute-wise transitions (comparing self and other merit or self and other outcome)) indicates the frequency of comparison types, with a Payne Index of 1 indicating only option-wise comparisons and a Payne Index of -1 indicating only attribute-wise comparisons. We find consistently negative Payne Indices across experiments and decision types: Free Focus Involved = -0.43; Constrained Focus Involved = -0.45; Free Focus Impartial = -0.49; Constrained Focus Impartial = -0.53, supporting a focus on attributes in the analyses.

Table 1.3: Overview of the main treatment effects

	Hypothesis 1 % given to Adv.		Hypothesis 2 Δ Attention		Hypothesis 3 % given to Adv.	
	Involved (1)	Impartial (2)	Involved (3)	Impartial (4)	Involved (5)	Impartial (6)
Advantaged	10.0***	3.44***	-0.15*	-0.21*		
	(1.00)	(0.70)	(0.074)	(0.10)		
Outcome Focus					2.93*	0.82
					(1.36)	(0.84)
Experiments	Constrained Focus and Free Focus				Constrained Focus only	
Observations	11930	11923	11930	11923	7942	7939

All models are linear regressions. Data: Columns 1, 3, and 5, Involved trials; Columns 2, 4, and 6, Impartial trials; Columns 5 and 6 exclude the dictators from the Free Focus treatment. Dependent variables: in Columns 1, 2, 5, and 6, the percentage of the pie allocated to the Advantaged member of the pair; in Columns 3 and 4: difference in dwell time between merit and outcome information. Standard errors clustered by participant in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. List of controls common to all regressions: age, gender (man, woman, other), political affiliation (5 categories), education (6 categories), income (7 categories), continent (4 categories). In addition, Columns 1, 2, 5, and 6 include the share of correct answers coming from the advantaged member over the total number of correct answers of the pair, task type (4 categories).

all our regressions focused on allocation decisions. Online Appendix A.1.5 includes a figure of the relationship between the share of correct answers and allocations, split by status. Columns (3) and (4) of Table A.1.4 show that the share of the pie Advantaged dictators receive strongly and significantly increases with their share of correct answers.

Column (2) shows that allocation differences by Status persist into the Impartial trials, with the Advantaged dictators allocating significantly more to the Advantaged members of the pair. The differences in Impartial allocations are statistically significant but quantitatively smaller than in the Involved trials, accounting for less than half of the status bias. Combined, these results replicate prior work on behavioral allocation biases whereby participants randomly assigned a higher pay rate keep more for themselves (Konow, 2000; Rodriguez-Lara and Moreno-Garrido, 2012; Deffains, Espinosa and Thöni, 2016).

Hypothesis 2: Selective attention. Table 1.3, Column (3) regresses Δ Attention on a dummy for the Advantaged. It shows that across attention treatments, Advantaged dictators have lower Δ Attention ($p = 0.048$). That is, they pay relatively less attention to merit (and more to outcome) than Disadvantaged dictators in both Involved and Impartial allocations ($p = 0.038$). We thus observe selective attention (Hypothesis 2), whereby Advantaged dictators prefer information on performance that includes their artificial advantage. Section 1.4.4 shows that this treatment effect in entirely driven by the Advantaged dictators. Column (4) shows that attentional patterns formed in the Involved condition spill over into the Impartial trials.

Hypothesis 3: Attention impacts allocations. To investigate the causal impact of attention, we compare allocations to the Advantaged in the Outcome and Merit Focus treatments. Partic-

ipants gave 53.6% of the surplus to the Advantaged members of the pair in the Merit Focus treatment compared to 56.3% in the Outcome Focus treatment, a significant difference (rank-sum test $p = 0.028$). Columns (5) and (6) of Table 1.3 includes data only from the Constrained Focus treatments (Experiment 2), and show regression of the share of allocation to the Advantaged on a dummy for the Outcome Focus treatments. The coefficient for the dummy indicates that Advantaged members receive 2.93 percentage points more in the Outcome Focus treatment ($p = 0.033$). Thus, in line with Hypothesis 3, attention plays a causal role in allocations.

In the Impartial trials where the dictator's own payoff is not at stake, the difference between Outcome Focus and Merit Focus on allocations is smaller than in the Involved trials. Participants in the Merit Focus treatment, gave 53.5% of the surplus to the Advantaged members of the pair compared with 54.2% in the Outcome Focus treatment (rank-sum test $p = 0.33$; Table 1.3 Column (6)).

In short, we find evidence for all our main hypotheses. Below we discuss the determinants of attention and allocations in more detail, and investigate interaction effects between our treatment dimensions.

1.4.3 Determinants of Attention

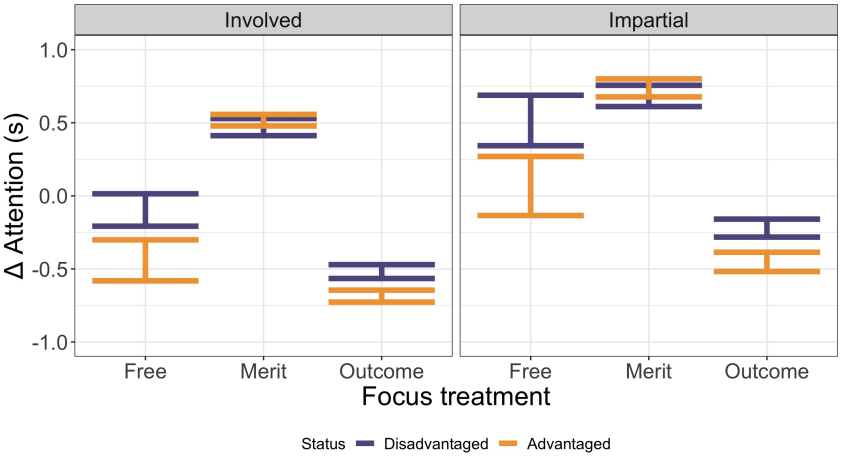
We now investigate how attention varies across all of our six treatments. Figure 1.5a) provides visual evidence of the average level of Δ Attention across our treatments. Table 1.4 provides corresponding statistical evidence in the form of OLS regressions, where we regress our main outcome variables on the treatment dummies. Columns (1) and (2) have Δ Attention as an outcome variable, whereas Columns (3) and (4) focus on allocations (see next section). Since the table does not include a constant, the coefficients for the three attention treatments represent the baseline levels of the attention and allocation variables for the Disadvantaged dictators. The interactions terms with Advantaged dummy show the change in Δ Attention for the Advantaged dictators.[13]

We first look at the Free Focus treatment, arguably the best test for selective attention, as it did not feature any restrictions on attention. The left panel of Figure 1.5a shows that Advantaged dictators spent about 350 ms longer on outcome information than Disadvantaged dictators, resulting in a more negative Δ Attention in Involved trials (rank-sum test of average dwell time $p = 0.011$). Column (1) of Table 1.4 mirrors this result with marginal statistical significance ($p < 0.1$). This difference in attention by Status is a result of the attention patterns diverging over time. Column (1) of Online Appendix Table A.1.6 regresses Δ Attention in the Involved rounds of the Free Focus treatment on the participant Status, the round number, and the interaction between Status and round number. There is no significant difference in Δ Attention for the

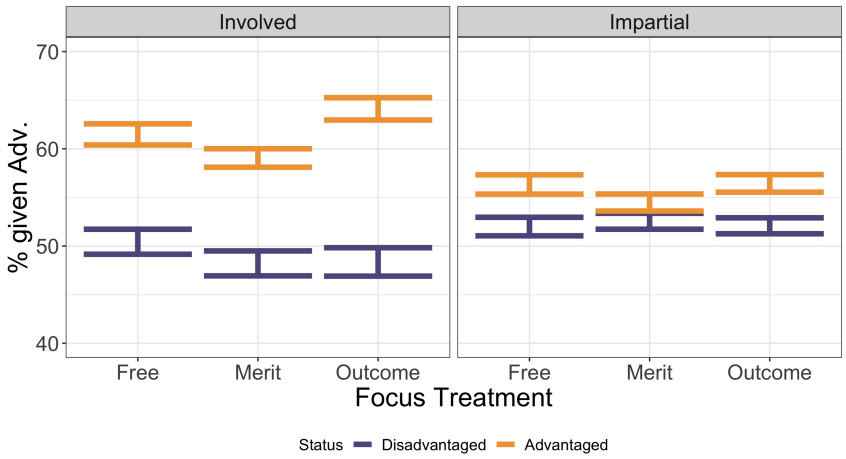
[13]Table 1.4 deviates from the preregistered analysis in not including the demographic and task type controls. We made this deviations to make the coefficients for "Free Focus", "Merit Focus", and "Outcome Focus" easier to interpret. In the current specification, these coefficients give the average value of the dependent variable for the Disadvantaged Dictators in these treatment. All the results presented in the Table replicate if we include the controls, as Online Appendix Table A.1.4 shows.

Figure 1.5: Overview of treatment effects on attention and allocations.

(a) Attention



(b) Allocations



The effect of Merit and Outcome Focus and Status on allocations and attention, shown separately for involved and impartial trials. The error bars represent the standard error.

Table 1.4: Interactions between status and attention treatments

	Δ Attention		% given to Adv.	
	Involved	Impartial	Involved	Impartial
	(1)	(2)	(3)	(4)
Free Focus	-0.093	0.51	50.4	52.0
	(0.11)	(0.17)	(1.27)	(0.95)
Free Focus * Adv.	-0.35^{+}	-0.44^{+}	11.0***	4.33**
	(0.18)	(0.27)	(1.67)	(1.37)
Merit Focus	0.47	0.69	48.2	52.5
	(0.058)	(0.072)	(1.27)	(0.81)
Merit Focus * Adv.	0.047	0.055	10.8***	1.93
	(0.070)	(0.095)	(1.58)	(1.19)
Outcome Focus	-0.52	-0.22	48.4	52.1
	(0.047)	(0.061)	(1.45)	(0.82)
Outcome Focus * Adv.	-0.17**	-0.23**	15.8***	4.38***
	(0.063)	(0.090)	(1.85)	(1.22)
Observations	11930	11923	11930	11923

All models are linear regressions. The models do **not** include a constant. Data from all dictators, Involved trials in Columns 1 and 3, and Impartial trials in Columns 2 and 4. Dependent variable in Columns 1 and 2: difference in dwell time between merit and outcome information. Dependent variable Columns 3 and 4: the percentage of the pie allocated to the Advantaged member of the pair. Standard errors clustered by participant in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

first round ($p = 0.72$), but Disadvantaged dictators pay relatively more attention to the merit information as the rounds progress ($p = 0.039$).

Turning to the Impartial decisions, the right panel of Figure 1.5a shows that the effect of Status on attention persists into the Impartial trials, although with larger variance (rank-sum test $p = 0.044$). Table 1.4, Column (2) shows that this effect is significant at the 10% level ($p = 0.094$), but not in regressions with demographic controls as shown in Online Appendix Table A.1.4. These mixed results in the Impartial trials may be due to the fact that attention was more variable (higher standard deviations in Column (2) than in Column (1) of Table 1.4) and information avoidance (see subsection 1.5.5) was higher than in the Involved trials. These two facts suggest that participants may have cared less about the information when their own payoff was not at stake.

Turning to the Merit Focus and Outcome Focus treatment, we confirm that the constraints in these treatments were effective in actually shifted attention. Table 1.2 shows that the percentage of time spent looking at merit information was 47% in the Free Focus treatment, that the Merit Focus treatment increased the value to 60%, whereas the Outcome Focus treatment decreased the value to 43%. This translates to a difference between Merit Focus and Outcome Focus in Δ Attention of around 1 second in both Involved and Impartial trials as shown in Figure 1.5a (rank-sum tests $p < 0.001$). The impact of Status on attention is present in the Outcome Focus treatment (rank-sum test $p = 0.011$) but not the Merit Focus treatment (rank-sum test $p = 0.45$). Column (1) of Table 1.4 replicates these results with a regression. given the constraints we imposed, it is perhaps not surprising that the Status differences in attention are less pronounced.

Finally, we note that both Advantaged and Disadvantaged participants spent relatively more time on merit information in the Impartial trials. While we did not hypothesize this pattern, it is consistent across experiments and suggests that merit information was considered relatively more important in the absence of self-interest motives.

1.4.4 Determinants of Allocations

We turn to determinants of allocations, as measured by the the share given to the Advantaged dictator. Average allocations across all six treatments are illustrated in Figure 1.5b and in Columns (3) and (4) of Table 1.4.

Above, we have already established the causal effect of Status and the Outcome Focus treatment on the share given to Advantaged. Figure 1.5 disaggregates this result. In the Involved trials, the effect of Advantaged is robust across all three attention treatments (rank-sum test $p < 0.001$ in each case). Indeed, in Table 1.4, Column (3), the interaction term for Advantaged is highly significant in each treatment. However, the size of the status effect fluctuates: it is lowest in the Merit Focus treatment at 10.8 percentage points and highest in the Outcome Focus treatment at 15.8 percentage points. This suggests that the difference in allocations between the Constrained Focus treatments documented in Table 1.3 is driven by the Advantaged dictators, who shift their allocations between Merit and Outcome Focus by almost 5 percentage points (or 0.58 of a standard deviation) - a substantial effect also compared to the 0.2 percentage point shift

by Disadvantaged dictators. Online Appendix Table A.1.7, Column (1) formally confirms that there is a positive interaction between Status and the Focus treatments on allocations significant at the 10% level ($p < 0.067$). Section 1.5.1 introduces a theoretical model that can capture this interaction and discusses the intuition behind it.

In the Impartial trials, splitting up the effect of Constrained Focus treatments by Status shows a similar pattern. Table 1.4, Column (4) shows that Status differences in allocations persist into the Outcome Focus treatment, but go away in the Merit Focus treatment. These findings mirrors the effects of attention documented above, and are in line with the idea that Advantaged dictators struggle to justify their higher share when they are forced to focus on the merit information.

Quantifying the impact of dwell time. To get a better sense of the quantitative importance of dwell time, we investigate how increasing Δ Attention by a given amount, say one second, affects allocations. A one-second increase (reallocating 500 ms from merit to outcome information) implies a shift equivalent to 23% of the average dwell time in the Constrained Focus treatment. Such an increase in Δ attention is similar to the one produced by the Outcome Focus treatment, so it does not involve an extrapolation of our treatment effects.

Our analysis is based on a 2-stage instrumental variable regression, where we instrument dwell time with the Focus treatment to which the subject is assigned, pooling the data at the subject level.[14]. Table A.1.3 shows the result of the second stage regressions. Column (1) shows that increasing Δ Attention by one second leads to a 2.6 percentage point decrease in allocations to the Advantaged members. Moreover, to compute the impact of a one-second change in Δ Attention on the effect of Status, we repeat the IV analysis separately for the Advantaged and Disadvantaged dictators in Columns (2) and (3) of Table A.1.3. Increasing Δ Attention by one second cuts the share that the Advantaged keep for themselves by 4.1 percentage points ($p < 0.001$), whereas it cuts the share that Disadvantaged dictators give to Advantaged recipients only by 0.1 percentage points, a negligible and insignificant effect. Thus, changing Δ Attention by one second reduces the gap between the allocation of Advantaged and Disadvantaged dictators by 4 percentage points ($p = 0.087$).[15] We conclude that reallocating 500 ms (or 23%) of dwell time from Outcome to Merit information reduces the effect of Status on Allocation in the Free Focus treatment by 36%.

[14] The F-statistic of our first stage is above 550, indicating a strong instrument and a minimal expected bias in the estimates. The exclusion restriction – that attention constraints only affect allocations via dwell time – is in line with standard models of attention like drift diffusion models, which focus on dwell time as the exclusive variable (Krajbich, Armel and Rangel, 2010). We can also exclude that our restrictions have a demand effect - see Section 1.5.3. Furthermore, the time limit on at least one box is binding in 90.5% of the Involved trials, indicating that our IV estimate is informative about most of our observations. Furthermore, the monotonicity assumption (Imbens and Angrist, 1994) is satisfied in our setting because would-be defiers have no way to alter the time restrictions on a box in a given round. Pooling the data at the individual level is necessary because the instrument - the Focus treatment - varies between but not within subjects. As such, in the second stage, a participants' predicted Δ Attention is the same in every round.

[15] To obtain this p-value, we run an IV regression with all the data from the Constrained Focus treatments. In it, we included Δ Attention and its interaction term with Status and we used the Outcome Focus treatment and the interaction between the Status and Outcome Focus treatments as instruments. We then test whether the interaction term is different from zero. Column (3) of Table A.1.7 in the Online Appendix reports this IV estimation.

Panel B of Table A.1.3 reports the corresponding results for the Impartial trials. Compared to the Involved trials, the results go in the same direction, but with a smaller effect size and less statistical significance. In particular, Column (3) shows that increasing Δ Attention by one second cuts the share that the Advantaged dictators give to the Advantaged recipients by 1.5 percentage points, a marginally significant difference ($p = 0.093$), and that the effect of Status on allocation goes down by 2.4 percentage points or 54% of the effect of Status on allocation found in Column (4) of Panel B of Table 1.4 ($p = 0.077$). Thus, shifting less than a quarter of the attention can eliminate more than half of the self-serving biases in allocation, as measured by the effect of status on Impartial allocations (Konow, 2000). This is a large effect and future research should investigate the robustness of this result.

How much does endogenous attention change allocation decisions? Here, we use the results of the Constrained Focus treatments to estimate the impact of voluntary changes in attention in the Free Focus treatments. We focus on Advantaged dictators, as they are most affected by the shifts in attention. We perform a simple back of the envelope calculation, using the fact that Advantaged dictators keep 4.1 percentage points more in the Involved trials if Δ Attention increases by one second (Table A.1.3, Column (3)). Moreover, from Table 1.3, we know that Δ Attention is 0.35 seconds lower for the Advantaged dictators than for Disadvantaged dictators in the Free Focus treatment. Multiplying these two numbers, we predict that the endogenous shift in Δ Attention in the Free Focus treatments causes the Advantaged dictators to keep 1.43 percentage points less of the surplus. This drop is equal to 13% of the difference in allocations between Advantaged and Disadvantaged dictators in the Involved allocations of the Free Focus treatment. If we repeat the same calculations for the Impartial Allocations, we find that the Advantaged dictators would have given 0.68 percentage points (or 16%) less to the Advantaged member of the pair if they had looked at the information as the Disadvantaged dictators did.

Of course these are crude calculations, as they assume that the effect of Δ Attention on behavior is linear and equally large across the different focus treatments. Nevertheless, they suggest a sizable impact of selective attention on behavior.

1.5 Discussion

In this section, we discuss the interpretation of our results. First, we introduce a theoretical model to guide the interpretation of our results. Then, we show the impact of attention on adherence to fairness criteria. Finally, we discuss and rule out potential confounds including experimenter demand effects and processing errors, and discuss other attentional measures.

1.5.1 Theoretical interpretation

To further guide our interpretation of our results, we provide a theoretical model in Appendix A.2. Building on Konow (2000) and (Cappelen et al., 2007), we assume dictators feel guilty about keeping more than their fair share, determined by subjectively applying fairness criteria.

We introduce attention into this framework, and we assume that paying attention to a fairness criterion increases its subjective weight. Thus, dictators' attention reflects a trade-off between attending to the criterion that justifies keeping most of the money and a psychological cost of distorting attention.

This model can generate our main hypotheses. It turns out that the model can also explain the observed, but not hypothesized, asymmetry between Advantaged and Disadvantaged dictators under some mild additional assumptions. The intuition for this result comes from the way status affects the optimality of different fairness criteria, and is similar to that in Hochleitner (2022). The egalitarian split is relatively better for the Disadvantaged dictators because the egalitarian criterion is either the criterion that gives them the most or it is at least the second-best criterion for them. Vice versa, the egalitarian criterion is never the best criterion for the Advantaged dictators.[16] Hence, the Disadvantaged dictators are better off placing a higher subjective weight on the egalitarian criterion than the Advantaged ones are. Since the egalitarian split can be achieved without paying attention to any performance information (except the total surplus), this makes Disadvantaged dictators' decisions somewhat inelastic to attentional shifts. By contrast, the Advantaged dictators reduce guilt by placing a high weight on the appropriateness of using raw outcomes (a "Libertarian" criterion - see also the next section). Doing so requires them to spend enough time on outcome information, making them relatively responsive to attentional constraints in this dimension.

1.5.2 Does Attention Affect Perceptions of Fairness?

One way in which attention may change behavior is through the perception or internalization of normative fairness criteria. For instance, participants for whom merit information is available relatively longer may be more likely to consider this information ethically relevant for their allocation. To investigate this mechanism, we look at dictator adherence to three criteria that are often invoked in the fairness literature (e.g. Konow, 2000; Cappelen et al., 2007; Bortolotti et al., 2017). The *Egalitarian* criterion requires splitting the surplus in equal parts among participants. The *Meritocratic* criterion requires splitting the surplus proportionally to the ratio of correct answers of the two participants in the real effort task. Finally, the *Libertarian* criterion requires splitting the surplus proportionally to the ratio of monetary contributions of each participant in the pair. The latter two criteria depend explicitly on information about the task performance of both participants in the pair, whereas the Egalitarian criterion can be implemented in the absence of any information.

Our main measure of fairness perceptions are the dictator allocations in the Impartial trials, which eliminate considerations of personal gain. We consider an allocation to be consistent with a fairness criterion if the distance between the chosen allocation and the prescription implied by the criterion is less than 5% of the total surplus size. Defined in this way, 20% of the choices are Egalitarian, 35% are Meritocratic, and 23% are Libertarian.[17]

[16] This statement is true if the Disadvantaged don't answer many more questions correctly than the Advantaged,

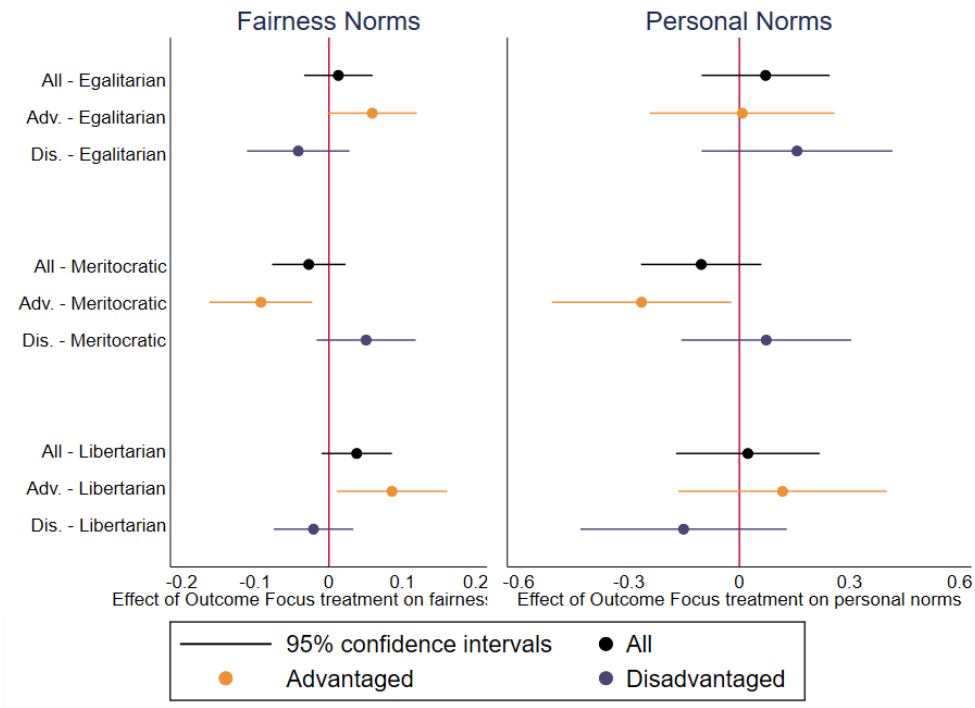


Figure 1.6: The effect of the Outcome Focus treatment of Fairness choices (left panel) and Personal Norms (right panel). For the left panel, the dependent variable is adherence to fairness criteria in dictator allocations in the Impartial trials. Pictured effects represent coefficients of dummy of Advantaged in linear regression models. The 95% confidence intervals are computed using standard errors clustered at the dictator level. For the right Panel, the dependent variable is the dictators' ratings about the moral appropriateness of redistributing according to the different fairness criteria. Pictured effects represent coefficients in an ordered logit regression. Pictured effects represent coefficients of dummy of Advantaged in ordered logit models. The 95% confidence intervals are computed using robust standard errors.

To investigate the impact of attention on this fairness measure, we define a dummy that takes a value of 1 if the impartial allocation adheres to the relevant fairness criteria, and regress this on the Outcome Focus treatment in the Constrained Focus treatments. Because we found above that Advantaged dictators are more susceptible to our Outcome Focus treatments, we show the aggregate effect as well as the effect split by Status.

The left panel of Figure 1.6 shows the results of this analysis. The Outcome Focus treatment (relative to the Merit Focus treatment) causes a modest shift towards more Libertarian choices and away from Meritocratic choices. The regression results underlying this graph are given in Online Appendix Table A.1.13. They show that while aggregate shifts are not statistically significant, the shifts for the Advantaged dictators are. The shifts for the Egalitarian criterion are not statistically significant at the 5% level for either group of dictators.

To further investigate these patterns, we look at a secondary measure of fairness, namely dictators' ratings of "moral appropriateness" of the different fairness norms, measured on a Likert Scale from 1 to 4. The right panel of Figure 1.6 shows the coefficient of ordered logit regressions, where the dependent variable is the dictator endorsement of the relevant fairness criteria (see Online Appendix Table A.1.14 for the associated regressions). The results follow the same pattern as those of the fairness allocations with the Outcome Focus treatment leading to a modest shift towards Libertarian norms and away from Meritocratic norms for Advantaged dictators. However, these results are noisy and statistically significant only for the shift away from Meritocratic norms.[18]

An interesting further question is how the status manipulation affects allocations and fairness views in the Impartial trials. Table A.1.12 in the Online Appendix shows evidence that Advantaged dictators are about 6 percentage points less likely to make Egalitarian split and about 10 percentage points more likely to make a Libertarian split. These estimates support the idea that dictators adopt self-serving views of fairness. For reasons of space, we leave a detailed examination of this effect and the relation to our secondary elicitations of fairness views to a companion paper (Amasino, Pace and Weele, 2023).

1.5.3 Experimenter Demand Effects

During the design phase of the experiment, we worried that our attention manipulations might give subjects a feeling that some information was deemed more important, inducing experimenter-

a condition that is almost always satisfied in our experiment.

[17]For example, we consider any allocation for which a member of the pair receives between 45% and 55% of the surplus to be consistent with the Egalitarian criterion. Using these definitions, 78% of the allocations are consistent with at least one fairness criterion and 66% of the allocations are consistent with only one criterion. In some rounds different criteria require similar allocations. For example, this happens if the participants answered the same number of questions correctly in a task. In that case, both the egalitarian and the meritocratic criteria require an equal split.

[18]In addition, we asked for dictators expectations of other's endorsement of the same norms ("social appropriateness"), using the method by Krupka and Weber (2013). Figure A.1.8 in the Online Appendix presents an analysis of this variable, analogous to the left panel of Figure 1.6. We find similar asymmetries between the Advantaged and Disadvantaged dictators in the response to the Outcome Focus treatment, but effects are noisier and not significant. This may reflect measurement error as well as participants' uncertainty about how other participants evaluated fairness norms.

demand effects. To counter this and obfuscate the research goal, 6 of the 20 decision rounds featured attention manipulations that were orthogonal to that of the treatment, as described in Section 1.3.3. In addition, our questionnaire featured several questions about the perceived goal of the experiment and the perceived direction of the attention restrictions.

The final questionnaire clearly shows that demand effects are not an issue: on a free form question, *none* out of 400 dictators indicated that the box timing was a purpose of the experiment. Moreover, it appears our obfuscation strategy was successful: when asked explicitly whether they perceived a difference in the timing closing of boxes, 60% of participants said they did not detect a systematic difference in box closing times. Overall, only 20% guessed the restrictions on both boxes correctly, a further 5.5% guessed one box correctly, and 8.5% guessed entirely wrongly.

To see if demand effects may have played a role, we test Hypothesis 3 using the same regressions as before, but restricting our sample to the 60% of participants who did not detect any difference in closing time. Table A.1.15 in the Online Appendix provides the results of this analysis. We replicate our finding that attention changes allocation decisions. If anything, the results are *stronger* in this sub-sample. This further demonstrates that experimenter demand effects did not drive our results.

1.5.4 Dwell Time Restrictions and Processing Errors

Our attention treatments were designed to measure the impact of the length of time subjects engage with information, while preserving subjects' possibility to process each source of information. The attention recognition literature suggests that recognition and memory consolidation for more complex scenes only takes up to 400 ms, and other processing studies have used dwell times of 250 ms or mouselab box times of 300 ms (Potter, 1976; Potter et al., 2014; Armel, Beaumel and Rangel, 2008; Milosavljevic et al., 2012; Pärnamets et al., 2015; Pachur et al., 2018; Ghaffari and Fiedler, 2018; Fisher, 2021). Therefore, 400 ms is well-within the recognized time-window for processing a single piece of information.

In addition, there are several ways our data can identify potential processing errors. First, Disadvantaged dictators do not change their allocations with the different attention restrictions (see Tables A.1.3 and A.1.12). For instance, we do not see that the Outcome Focus treatment leads Disadvantaged subjects to adhere less to Meritocratic and more to Libertarian and/or Egalitarian criteria. This result shows that subjects are able to choose the same information-based allocations under any type of restriction and speaks against the restrictions a having a mechanical effect on allocation.

Second, we can exploit within-subject variation in dwell time restrictions. Recall that in every attention treatment, the attention restrictions on one type of information were implemented only in 14 out of the 20 rounds. In the remaining 6 rounds, the restrictions were randomly allocated to other dimensions (see Section 1.5.3). Thus, if the restrictions affected dictators' allocations through processing errors, we should see a difference between the 14 treatment-congruent trials with the 6 remaining trials. For instance, we should see that Advantaged subjects in the

Outcome Focus treatment are more generous in the remaining 6 trials where merit information was less restricted. Online Appendix Table A.1.16 shows the results of regressions that include trial-by-trial dummies of dwell time restrictions (Self, Other, Merit, Outcome), in addition to our main treatment dummy. We find that the type of trial has no statistically or quantitatively meaningful impact on behavior beyond our main treatment. Furthermore, the effect of the Outcome Focus treatment on allocation does not go down once we control for trial type. This shows attention in any single trial does not have a strong influence on behavior, but rather that it is the sustained push to attention over multiple trials that produces the effect of the focus treatment.

In summary, the attention manipulation did not prevent subjects from making any particular allocation. Of course, dwell times may affect the ease with which subjects can incorporate information into decisions, but this is exactly the point of studying this variable in the first place.

1.5.5 Information Avoidance

Our focus is on continuous measures of dwell time and relative attention, wherein participants have processed all relevant information, but simply place different weights according to the time spent on it. This is qualitatively different from a previous literature looking at binary information seeking or avoidance designs, where participants do not have access to the information they avoid. In such cases, information avoidance may signal that participants decide independently of merit or outcome and thus have no use for the information, or that they want to avoid information in order not to face psychological conflicts from taking the most money for themselves (Dana, Weber and Kuang, 2007; Grossman and Van der Weele, 2017).

We find that information avoidance does not play an important role in our experiment. Dictators open all the boxes in 84% of Involved trials, and avoidance of either type of information is lower than 10% on aggregate. In the Impartial trials, information avoidance is higher, but subjects still open all boxes in 70% of trials. In addition, there are no clear self-serving patterns in the avoidance behavior, as we discuss in Online Appendix A.1.15. To confirm that avoidance does not drive our results, we replicate all our findings excluding trials with avoidance and collapsing the data at the individual level. In the Online Appendix, Tables A.1.18 and A.1.19 show that all our results hold in this restricted data. Thus, it appears that selective attention occurs on the intensive, rather than the extensive margin.

1.5.6 Other results discussed in the Online Appendix

Dwell time is not the only measure of attention found to matter in choice. Other important measures in process-tracing include the instances of looking at information (i.e. the number of times each box is opened) and the last information examined (Willemsen and Johnson, 2019; Rahal and Fiedler, 2019). Online Appendix A.1.10 replicates our findings using these other measures.

1.6 Conclusion

In this paper, we show that economic advantage causes selective attention, as it reduces how long people dwell on information about merit. Furthermore, we demonstrate the causal impact of dwell time on behavior, as biased attention increases the amount of money people allocate to themselves or other similarly advantaged individuals. Some of these effects persist, albeit in somewhat weaker form, in situations where people have to make decisions between two other individuals and their own income is not at stake. In particular, we show that in such settings, attentional shifts cause more libertarian and fewer meritocratic allocations among Advantaged dictators. As underlying psychological mechanisms, we can rule out experimenter demand effects and processing errors, and find evidence that sustained attentional manipulation affect the formation of fairness views. We go beyond a previous literature on information avoidance, as we show that it is relative dwell time, and not the pure avoidance of information that drives our results.

Quantitatively, the effect of attention on decisions in the experiment is substantial, and reduces self-serving bias by a meaningful amount. This provides a promising base for further research on the design of interventions and policies based on visual attention, such as online information campaigns or educational campaigns to combat bias. It also suggests that political advertising about the sources of inequality on social media or elsewhere can affect attitudes for redistribution.

These results show the importance of attention to effort and luck for redistributive behavior. More research is needed to determine the ecological validity of these claims. Evidence on self-serving biases in the laboratory have been confirmed in natural experiments (Di Tella, Galiant and Schargrodsky, 2007; Hvidberg, Kreiner and Stantcheva, 2020; Schwardmann, Tripodi and Van der Weele, 2022), so future research could establish whether the same is true for the attention channel identified in this paper. Given the complex experimental design and multiple analyses, more research examining the relationships between status, attention, and allocations across different contexts are needed to confirm the robustness of our findings.

Extrapolating for a moment beyond the laboratory, selective attention may explain why groups have different views on the nature and desirability of inequality, and provide insights for a current debate about the role of meritocracy in Western society. For instance, elites' attentional habits may cement views that wealth differences are earned, explaining the findings of recent surveys. It can also explain why elites favor policies promoting open markets and low redistribution, while looking away from the institutionalized advantages that allow them to reap disproportionate benefits of such policies (Sandel, 2020). Future research could explicitly study the media consumption of those groups, and test whether exposing people to different types of information helps to reduce polarization in beliefs outside the laboratory. The results could be relevant for other domains where a subgroup of society enjoys institutionalized advantages, whether they are based on income, race or gender.

Chapter 2

Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms

This chapter is based on Amasino, Pace and Weele (2023). The results of this chapter are based on the same experiment and on the same data used in Chapter 1.

2.1 Introduction

People with higher incomes often support less redistribution than those with lower incomes, a finding that has been consistently shown across surveys, field, and lab experiments (Koo, Piff and Shariff, 2022; Suhay, Klašnja and Rivero, 2021; Cohn et al., 2019; Konow, 2000; Di Tella, Galiant and Schargrodsky, 2007). This gap in support for redistribution could be due purely to self-interest. However, in line with self-image and reputational motivations to appear moral to oneself or others, people often do not go to selfish extremes. Instead, they find excuses or justifications that allow them to support fairness ideals that most benefit themselves. This is especially pernicious in privileged or powerful individuals who are in a position to institutionalize their self-serving bias[1], which has been linked to polarization, resentment, and social conflict in Western democracies (Piketty, 2020; Sandel, 2020; Babcock et al., 1995; Schwardmann, Tripodi and Van der Weele, 2022).

While self-serving redistribution decisions are well-documented, their psychological antecedents are less well-understood. Theories of fairness and cognitive dissonance have invoked various psychological pathways, including shifts in personal fairness views (Konow, 2000), biased perceptions of social norms (Bicchieri, Dimant and Sonderegger, 2023), or motivated beliefs about merit and returns to effort (Bénabou and Tirole, 2006; Deffains, Espinosa and Thöni, 2016). Many empirical papers have looked at the role of individual psychological constructs, but

[1]Note: our definition of self-serving bias is self-serving judgments of a fair division (Rodriguez-Lara and Moreno-Garrido, 2012; Cappelen et al., 2007). These self-serving biases are different than the common definition of self-serving attribution bias in social psychology, which means to attribute good outcomes to one's ability or effort while attributing bad outcomes to external circumstances such as bad luck (Deffains, Espinosa and Thöni, 2016; Dorin et al., 2021; Miller and Ross, 1975; Bradley, 1978).

there are few comparisons of their relative importance. Moreover, error in the measurement of these constructs has complicated the effort to quantify their explanatory power.

In this paper, we directly measure and investigate the role of these three psychological constructs in redistribution decisions, examining how each construct is affected by status and its potential mediating role in the effect of status on redistribution decisions. First, we look at "personal norms" that characterize what people regard as fair. Personal norms reflect privately held views of fairness that develop out of experience and moral reasoning. They are predictive of pro-social or selfish behavior in economic allocation decisions (Bašić and Verrina, 2021; Messick and Sentis, 1979). Second, we look at "social norms", that is, people's perceptions of what others think is fair. In our setting, social norms are determined by beliefs about which fairness principle(s) most people endorse. The desire to conform with others' views makes social norms predictive of individuals' actions (Krupka and Weber, 2013).

While personal and social norms often align, they are different constructs and can diverge in meaningful ways. For example, most young, married men in Saudi Arabia privately support women working outside the home. Still, a presumed social norm against women's labor force participation undermines support for their wives' job searches (Bursztyn, González and Yanagizawa-Drott, 2020). In the context of climate change, Sparkman, Geiger and Weber (2022) and Andre et al. (2021) find that most Americans are willing to support mitigation efforts, but they underestimate others' support for mitigation, undermining collective action. Findings from experimental data on allocation decisions also suggest that these constructs have separate predictive power for behavior (Bašić and Verrina, 2021). Moreover, the extent to which different constructs predict behavior may depend on the strength of social image concerns and expectations of conformity (Bašić and Verrina, 2021; Thøgersen, 2008; Ajzen and Fishbein, 1970; Cialdini, Kallgren and Reno, 1991).

Third, we consider beliefs about the determinants of economic success and inequalities. High-income people are more likely to attribute their success to hard work and ability than luck (Suhay, Klašnja and Rivero, 2021; Valero, 2021; Deffains, Espinosa and Thöni, 2016; Dorin et al., 2021; Cassar and Klein, 2019; Di Tella, Galiant and Schargrodsky, 2007). In contrast, those who are less successful or experience hardship are more likely to point to the role of luck or selfishness in success (Hvidberg, Kreiner and Stantcheva, 2020; Hochleitner, 2022; Almås et al., 2022). Beliefs about the determinants of success have been shown to influence people's preferences for redistribution, as people are more likely to redress inequalities due to luck rather than differences in effort (e.g. Cherry, Frykblom and Shogren, 2002; Krawczyk, 2010; Cappelen et al., 2013; Durante, Putterman and Van der Weele, 2014; Lefgren, Sims and Stoddard, 2016; Cappelen et al., 2017; Bortolotti et al., 2017).

In this study, we investigate with an experiment how having a privileged status impacts these three constructs, and we study their mediating role in allocation decisions. We do so in the context of a large online experiment with a sample of 600 participants based on the design of Konow (2000). In the experiment, participants first work on real-effort tasks to produce earnings. We manipulate status by randomly assigning half of the participants a higher pay rate per correct answer in the tasks, such that half have a pay advantage and half have a pay

disadvantage. Participants then act as "dictators" deciding how to divide joint task earnings, first between themselves and another participant and then between two others in which they have no stake of their own. In this setting, we replicate the findings of Konow (2000), who showed that participants advantaged by a randomly-assigned higher pay rate keep more of the joint earnings and continue to favor other advantaged workers even when self-interest is removed. This persistence to "impartial" decisions is particularly indicative of self-serving bias, and it is the focus of our investigation.

Our original contribution is in (1) examining how the randomly-assigned (dis)advantage in pay rate (or 'status') impacts personal norms, social norms, and beliefs and (2) quantifying and comparing the mediating roles of each construct in the relationship between status and divisions of joint earnings accounting for measurement error. We find that status differences lead to self-serving shifts in personal norms and beliefs, but we find no statistically significant effect for social norms. Moreover, participants show awareness of the bias induced by status in fairness principles when predicting others' norms. Finally, we show that differences in divisions of joint earnings due to dis(advantaged) status in impartial decisions are primarily mediated by shifts in personal norms, with minimal contributions of social norms and beliefs. This result points to a primary role of shifting personal norms (without significant changes in perceptions about what others find appropriate) in driving self-serving attitudes toward redistribution.

Our findings go beyond existing empirical work that either infers psychological mechanisms from shifts in behavior or focuses on a particular mechanism. Konow (2000) and Rodriguez-Lara and Moreno-Garrido (2012) found that participants who benefit from luck incorporate it into their fairness principle when dividing joint earnings, supporting the idea that personal norms adapt to the context. However, they do not explicitly measure personal norms, social norms, or beliefs – they infer this from allocation choices. Deffains, Espinosa and Thöni (2016) identify self-serving biases in the selection of redistribution criteria as well as a corresponding shift in attribution whereby more successful dictators are more likely to attribute their success to effort. However, they do not explicitly study the link between these variables. Dorin et al. (2021) use the setup of Deffains, Espinosa and Thöni (2016) to explore the role of in-group bias and personal norms as mediators of self-serving biases, finding that both act as contributing mechanisms of the bias. Valero (2021) and Lobeck (2021) show that participants distort beliefs about performance independently of monetary incentives to do so. Yet, they do not quantify the mediating role of beliefs in self-serving biases. Ubeda (2014) runs a descriptive study where she classifies the dictators' fairness norms.

2.2 Theoretical framework

Our introduction cites work showing that socio-economic status affects beliefs about fairness and merit and attitudes towards redistribution. To explain these observations, several papers have invoked concepts like cognitive dissonance (Konow, 2000) or motivated reasoning (Suhay, Klašnja and Rivero, 2021). According to such accounts, the wish to justify the status quo and

limit redistribution to the less fortunate leads people to self-servingly manipulate their fairness ideals and attributions of success. In Online Appendix B.2, we formalize this idea in a model inspired by (Cappelen et al., 2007). The model captures a simple division problem – mirroring the setup of the current experiment and earlier experiments – where a decision maker allocates a sum of money that has been produced by herself and another person. Crucially, one of the two agents randomly receives a relative "advantage" in the production process, whereby her performance is multiplied by a higher pay rate, boosting her production share in the total surplus to be divided.

When dividing the surplus, we assume that decision-makers care both about their own payoff and about the fairness of the allocation. Specifically, we assume they adhere to one of several fairness criteria that have been identified in the literature (Konow, 2000; Cappelen et al., 2007; Rodriguez-Lara and Moreno-Garrido, 2012): egalitarian (equal split), meritocratic (proportional to task performance), and libertarian (proportional to the share of total surplus produced - i.e., including randomly determined pay rate advantage). As fairness is subjective, agents may differ in which fairness criterion they deem most appropriate, or they may put some weight on all criteria. If the chosen allocation differs from their subjective fairness ideal, decision-makers incur a psychological cost in terms of self-image or guilt.

Thus, decision-makers in the model navigate a trade-off between taking more money for themselves and remaining closer to their subjective fairness ideal. This trade-off generates pressure to shift their subjective fairness ideal in a self-serving direction to increase the amount they can allocate to themselves without increasing guilt. As an example, consider an advantaged subject in the role of dictator. Because of her advantage, she will typically outperform the recipient in terms of the total contribution, although not necessarily on the “raw" task performance. This implies that the libertarian fairness criterion will be the most advantageous, as it prescribes taking a high share for herself.

We expand the model to capture the cognitive channels responsible for such self-serving bias. We assume that decision-makers may shift their weights on the different fairness criteria, as a function of their advantaged status. They can do so by changing their personal and social norms as well as the attributions of success. In terms of our example, we assume the advantaged decision maker may convince herself a) that the libertarian criterion is the most appropriate one (personal norms), b) that this view is generally shared among other participants so that she would find support for her decisions by others (social norms), and c) that her relative performance is higher than it actually is, so that she is entitled to a bigger share. In the model, these processes will increase the weight on the libertarian criterion in her fairness views, and/or reduce her experienced guilt level when she allocates money according to this (self-serving) criterion.

While our model illustrates the broad idea behind self-serving bias, it leaves open many details about how exactly norms and beliefs map into behavior. Thus, our main contribution is in the empirical quantification of the relative importance of different channels underlying self-serving biases. Further research can use these findings to model different psychological mechanisms in more detail.

2.3 Design

In this paper, we report the results of two experiments. Each experiment happened over 2 days: on Day 1, participants completed real effort tasks to generate a surplus, and on Day 2, participants in the role of dictators divided the surplus. Figure 1.2 in the previous chapter displays the timeline shared by the two experiments.

For Experiment 1, we recruited 200 dictators and 300 recipients from Prolific.co. The data was collected between the 13th and 19th of July, 2020. For Experiment 2, we recruited 400 dictators and 600 recipients from Prolific.co[2]. The data was collected between the 23rd and 30th of November, 2020. These sample sizes of 100 participants per treatment were preregistered (see preregistrations at the following links: Experiment 1, Experiment 2) and larger than those of similar studies (Konow, 2000; Rodriguez-Lara and Moreno-Garrido, 2012; Cappelen et al., 2007). The final sample of 600 dictators has 43% Women and the average age of dictators is 25.24 (standard deviation 7.27). Across both experiments, we paid a completion fee of $\pounds 2.85$ for Day 1 and $\pounds 6.15$ for Day 2 plus an average bonus of around £3 per participant.

2.3.1 Day 1: Surplus Generation

On Day 1, participants completed 8 real effort tasks. There were 4 different types of tasks: moving sliders to a predetermined position, logic questions, counting the number of zeros in a table, and solving Raven's matrices. Each type of task was repeated twice. In every task, each correct answer earned a monetary reward. When completing the tasks, the participants did not know the exact monetary reward they would receive. However, they knew that they would randomly be assigned a high or low pay rate per correct answer, the amount of both pay rates, and that they would learn which pay rate applied to them at a later stage. The high pay rate was always 3 times the low pay rate, but pay rates were calibrated (based on pilot data) according to task type to result in an average surplus of $\pounds 3.5$ per task.

Similarly, the participants were aware that the high or low pay rate assignment would apply to all of their tasks. We checked the participants' understanding of the randomness and persistence of the pay rates with two comprehension questions that they had to answer correctly to continue with the experiment. Participants were also informed that they would be paired with other participants and that their earnings would go into a single common account but they did not know how this would be divided.

2.3.2 Day 2: Surplus Division

After the Day 1 surplus generation, we split participants into dictator and recipient roles. Only the dictators were invited to Day 2, which started one day after Day 1. Day 2 was divided into

[2]We had 16 additional Dictators that started the second day of the experiment but did not complete it. Of those 6 are Advantaged and 10 are Disadvantaged; a Fisher's exact test does not reveal a statistically significant difference in the probability of completing the experiment for these two groups ($p = 0.45$).
We recruited more recipients than dictators because in the Impartial trials the dictators split the amount generated by two recipients.

3 parts. In Part 1, dictators split earnings between themselves and recipients, termed “Involved" allocations. In Part 2, they divided the earnings between pairs of recipients, termed “Impartial" allocations. In Part 3, they answered questions about their strategies, beliefs, and perceptions of norms.

At the beginning of Day 2, dictators learned their pay rate per correct answer. We call participants who received the high pay rate “Advantaged", those with the low pay rate “Disadvantaged," and we refer to this difference as the “Status” treatment. Participants then received instructions for the Involved allocation task. The joint earnings of a pair in a task were merged into a common account, and the dictator chose how to allocate this common account between themselves and the paired recipient. Over 20 trials, the dictators were matched with different recipients, with one of the 8 tasks underlying the common account in each trial. All recipients were assigned the opposite pay rate of the dictator, thus implementing inequality in the pair. During each trial, dictators received information about the relative contributions to the common account (more on that below) and made their allocation decisions.

In the next part of Day 2, dictators made Impartial allocation decisions for two recipients. Just as in the Involved allocations, the Impartial allocations always included one Advantaged and one Disadvantaged recipient. Over 20 trials, dictators chose how to divide the common account produced by pairs of different recipients. Participants always completed the Involved trials before the Impartial trials in order to test whether self-serving biases developed in Involved decisions persisted into Impartial decisions, as in Konow (2000) and to prevent the reverse effects (Dengler-Roscher et al., 2018). Such carry-over effects are relevant outside of the lab because people typically first experience their own economic status and may develop biases dependent on that status before making more abstract, impartial decisions about fairness for others. To control for purely mechanical carry-over effects in allocation, the orientation of the slider changed for half of the participants and the slider orientation is included in regressions looking at the impact of Status on allocation.

Decisions were incentivized by implementing one of each dictator’s 40 decisions. The average surplus per pair of participants in each task was £6.99 in Experiment 1 and £7.10 in Experiment 2. These amounts are approximately 1.4 times the minimum hourly wage on Prolific, so the allocation decisions had reasonably high stakes. If the decision came from the Involved allocations, the dictator received a bonus payment equal to the amount they kept for themselves, and the recipient received the amount allocated to them. If the decision came from the Impartial allocations, the dictator received $\pounds 1$, and each of the two recipients received what the dictator allocated them.[3]

Attention measurements and differences between Experiment 1 and 2.

Before every decision, the dictators had 6 seconds to look at information about the way the money in the common account was generated. Both experiments were also designed to study

[3]We pre-assigned which type of trial (involved or impartial) would be relevant for payment, and which recipients would get the bonus to ensure that all dictators and recipients were paid a bonus based on a single allocation decision. Recipients could appear in multiple different dictators’ allocation decisions.

the role of visual attention to this information, as described in the companion paper (Amasino, Pace and van der Weele, 2021). Participants could reveal information about the number of correct answers each participant in the pair completed – merit information – as well as the monetary contribution incorporating the randomly-assigned pay rate – outcome information. This feature was implemented in MouselabWEB, so participants could reveal each piece of information by hovering their mouse cursor over the relevant labeled box (Willemsen and Johnson, 2019). Experiment 1 measured naturally occurring attention patterns with no restrictions, whereas Experiment 2 had design features to manipulate attention and investigate its causal role. In Experiment 2, there were restrictions on the length of time (400 or 1600 ms per information box) that participants could reveal either the number of correct answers or monetary contributions within the total 6 seconds to look at information, pushing them to look at one of the pieces of information longer. This attention manipulation is the only difference in the allocation decisions between Experiment 1 and Experiment 2.

In this paper, we do not analyze attention. Instead, we focus on additional measurements of norms and beliefs across experiments and attention treatments. To ensure the attention treatments do not drive the results, all the regressions in this paper control for these attention treatments [4]. Moreover, all attention treatments were designed such that participants in each condition could access information about merit and luck. We further rule out that attention might be driving our results in Online Appendices B.1.6 and B.1.7.

2.3.3 Perception measurement

In Part 3 of both experiments, after the Involved and Impartial allocation decisions, we asked dictators a series of questions about their strategy, their perceptions of various fairness criteria, and their beliefs about the performance of different types of participants in the real effort tasks. For most of these variables, we conducted multiple elicitations per participant, a fact that we will leverage in the analysis. The main questions were always asked in the same order, but within a type of question, we randomized the order in which fairness rules were rated (e.g. libertarian, meritocratic, or egalitarian).[5] Moreover, we elicited participants' demographics, including gender, country, political leaning, education, and income level.

[4] To additionally test the effect of attention, we examined the interactions between our attention treatments and norm measurements. We do not find strong interactions, so the impacts of Status on norm endorsement do not seem to be primarily driven by attention. We find that, in the merit focus treatment, Advantaged dictators are more likely to endorse personal meritocratic norms. In contrast, Disadvantaged dictators predict higher social endorsement of libertarian norms, a somewhat counterintuitive result.

[5] In experiment 2, which has some additional elicitations compared to experiment 1, the order of elicitations was as follows: we first asked beliefs about performance. Next, we asked about personal norms, first an open-ended question about criteria for division followed by specific questions about the appropriateness of each fairness criteria. After personal norms, we asked about overall social norms, followed by eliciting social norms specific to Advantaged or Disadvantaged dictators. Finally we asked another version of the personal norms questions about using only correct answers (merit) vs. only monetary contribution (outcome) to divide joint earnings.

Personal norms of fairness.

One channel for the development of self-serving biases is through the perception of what is morally appropriate behavior. In particular, Advantaged dictators may want to believe that inequalities due to luck are acceptable, while Disadvantaged ones might want to believe that these inequalities are unfair. We refer to people's fairness perceptions as "personal norms".

We obtained three independent measures of dictators' personal norms. Our main measure of personal norms is participants' ratings of the moral appropriateness of dividing according to three fairness criteria that are commonly used in the literature (Konow, 2000; Cappelen et al., 2007; Rodriguez-Lara and Moreno-Garrido, 2012): egalitarian (equal split), meritocratic (proportional to the share of correct answers), and libertarian (proportional to the share of total surplus produced - i.e. including randomly determined pay-rates).

Second, in Experiment 2 only, we asked participants to rate the moral appropriateness of allocating the surplus using different types of information. One question asked about the appropriateness of exclusively using the information about the number of correct answers, and the other about the appropriateness of exclusively using the information about the monetary contributions. While the framing is slightly different, the ratings from these questions map directly onto the appropriateness of different fairness norms. Specifically, using only information about the number of correct answers results in an allocation consistent with the meritocratic criterion, whereas using only the information about the monetary contributions results in a split consistent with the libertarian criterion.

Finally, we asked dictators an open-ended question about how they redistributed the money. Unaware of the research question, a research assistant classified whether a participant's answer referred to the egalitarian, meritocratic, or libertarian criteria. Below, we exploit the common variation in these different elicitations to address errors in the measurement of personal norms.

Social norms of fairness.

To understand whether participants believed that their personal norms were commonly shared, we elicited their perceptions of social norms of appropriateness related to these criteria. To do so, we used the incentivized method from Krupka and Weber (2013): participants could win a $£1$ reward by correctly predicting the modal response to the appropriateness question for each of the three fairness criteria.

As a further measure of social norms, we also asked participants to predict the modal answers separately for Advantaged and Disadvantaged dictators. These elicitations had two purposes. First, they served to elucidate whether people can anticipate self-serving status bias in others. Second, they help with measurement error in the mediation analysis of Section 2.4.3.

Beliefs about relative performance.

Self-serving biases also arise via the formation of motivated beliefs about relative performance (Valero, 2021). Shifts in beliefs about the role of merit may affect how people think about

inequality and which social norms are relevant to their decisions. In Experiment 2, we additionally elicited incentivized beliefs about two different perceptions of relative performance. To encourage them to think carefully about these questions, participants could earn a $\pounds 1$ bonus for a correct prediction for each case.

First, we asked dictators to report the number of trials in which the recipient outperformed them (i.e., the recipient had more correct answers). The number of correct answers is the prime criterion of merit in the experiment, so forming motivated beliefs about this topic could provide a powerful justification for keeping more of the surplus. Through our construction of the experiment rounds, dictators had a higher number of correct answers in exactly 50% of the rounds, so we can compare the answer to a baseline of 50% that participants observed in their allocation decisions.[6]

Second, we measured how participants evaluated the size of the advantage. Advantaged dictators could justify allocating larger amounts to themselves if they believe that the pay rate inequalities in the experiment are too small to make a difference in output. We elicited this belief by asking for the share of pairs in which Disadvantaged participants produced more output than the Advantaged participants. A higher share corresponds to belief in a smaller relative advantage for the Advantaged participants. We expected Advantaged participants to be more likely to believe that the treatment gap was small such that Disadvantaged participants contributed more on average, reflecting thoughts like: “The receivers I was matched with performed poorly despite having a fair chance to produce a big share of the pie, so I should be entitled keep a larger share."

2.4 Results

We first characterize the self-serving bias by investigating the effect of the Status treatment on dictator behavior. We then look at the causal effect of Status on personal norms, social norms, and beliefs. Finally, we look at the role of personal norms, social norms, and beliefs in explaining the self-serving bias.

Table 2.1 provides an overview of the means and standard deviations of the primary outcome variables.

2.4.1 Status and Allocations

We investigate whether our results replicate those of Konow (2000). Figure 2.1 displays the share of the surplus that dictators allocated to the Advantaged member of the pair, split by Status, and by Involved or Impartial allocation decisions.

[6]We matched dictators and recipients in such a way that dictators answered more questions correctly in 50% of the rounds. We did so to reduce the between-dictator variance in the production inputs for the common account. There is only 1 dictator for whom this matching was not possible and who had a higher number of correct answers only in 40% of the trials.

Table 2.1: Names and definitions of main variables

Variable	Label	Definition	Advantaged	Disadv.
Involved allocation	% given to Adv.	The % of the common account allocated to the Advantaged participant in self-relevant decisions.	61.6 (10.8)	49.0 (13.4)
Self allocation	% Kept	The % of the common account kept by the dictator in self-relevant decisions.	61.6 (10.8)	51.0 (13.4)
Impartial allocation	% given to Adv.	The % of the common account allocated to the Advantaged recipient in impartial, self-irrelevant decisions.	55.8 (9.3)	52.2 (8.6)
Libertarian personal norms	perLib	Moral appropriateness rating (1-4) of the libertarian criterion: dividing according to monetary contributions (merit and luck).	2.75 (0.95)	2.54 (0.93)
Meritocratic personal norms	perMer	Moral appropriateness rating (1-4) of the meritocratic criterion: dividing according to the number of correct answers (merit only).	3.19 (0.82)	3.27 (0.81)
Egalitarian personal norms	perEga	Moral appropriateness rating (1-4) of the egalitarian norms: dividing evenly (regardless of merit or luck).	2.48 (0.86)	2.66 (0.88)
Libertarian social norms	socLib	Social appropriateness rating (1-4) of the libertarian criterion: dividing according to monetary contributions (merit and luck).	2.90 (0.96)	2.74 (0.95)
Meritocratic social norms	socMer	Social appropriateness rating (1-4) of the meritocratic criterion: dividing according to the number of correct answers (merit only).	3.23 (0.77)	3.29 (0.77)
Egalitarian social norms	socEga	Social appropriateness rating (1-4) of the egalitarian criterion: dividing evenly (regardless of merit or luck).	2.58 (0.86)	2.63 (0.88)
Recipient outperforming	# RecOutperf	Beliefs about the # of involved rounds (out of 20) experienced by the dictator in which the recipient had more correct answers.	8.10 (3.11)	9.68 (3.55)
Disadvantaged outcontributing	% DisOutcont	Beliefs about the % of rounds in which any Disadvantaged participant had a higher monetary contribution than an Advantaged participant.	20.95 (18.91)	25.06 (22.04)

Advantaged and Disadvantaged columns show the mean and standard deviation for each variable from both Experiment 1 and 2 for allocations and norms and from Experiment 2 only for beliefs.

Involved Allocations.

Focusing on the Involved allocations, a rank-sum test of the average share each dictator gave to the Advantaged members across rounds confirms that the two groups allocated significantly differently ($p < 0.001$). We confirm this result in regression analyses with standard errors clustered at the individual level and controls for subject characteristics, including gender, political orientation, and geographical background. Table 2.2, Column 1 provides the results of these regressions and shows that Advantaged dictators gave 10 percentage points more of the surplus to the Advantaged member (i.e., themselves) than Disadvantaged dictators gave to Advantaged recipients ($p < 0.001$), an effect that is almost as large as the standard deviation of allocation decisions for this group.

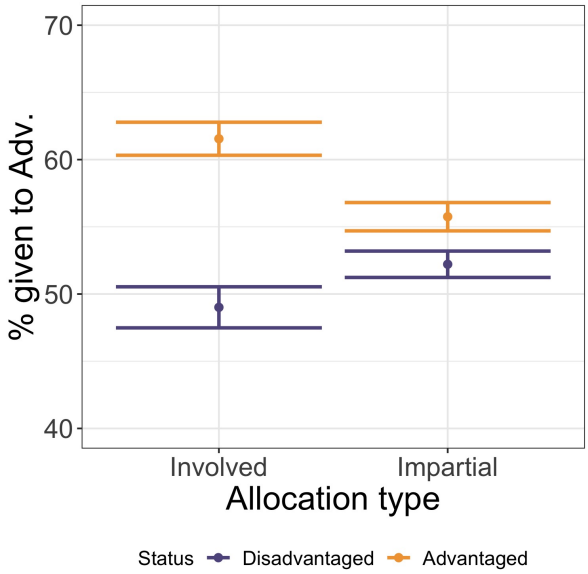
The fact that Advantaged dictators allocated more to themselves than Disadvantaged dictators allocated to Advantaged recipients is consistent with dictators simply keeping most of the surplus. Therefore, we look at the impact of the Status treatment on the share dictators kept for *themselves*. We see a very similar effect, with Advantaged dictators keeping 61.6% compared to Disadvantaged dictators keeping 51% (Table 2.1). This result is highly significant in both a rank-sum test ($p < 0.001$) as well as in a regression with controls (Table 2.2 - Column 2), and it replicates prior work on behavioral allocation biases whereby the participants randomly assigned a higher pay rate kept more for themselves (Konow, 2000; Rodriguez-Lara and Moreno-Garrido, 2012; Deffains, Espinosa and Thöni, 2016). In fact, the two ways of looking at the division are almost equivalent because the Disadvantaged dictators are very close to splitting the surplus 50-50 on average. This relatively even division accords with previous work showing that dictators respect earned income in their allocations (Cappelen, Sørensen and Tungodden, 2010; Rodriguez-Lara and Moreno-Garrido, 2012; Cherry, Frykblom and Shogren, 2002).

Table 2.2: Effect of Status on allocation

	(1)	(2)	(3)
	% given to Adv.	% Kept	% given to Adv.
Advantaged	10.0***	10.4***	3.44***
	(0.99)	(1.02)	(0.70)
Observations	11930	11930	11923
Trial type	Involved	Involved	Impartial

All models are linear regressions. Data from Experiments 1 and 2: Involved trials in Columns (1) and (2); Impartial trials in Column (3). Dependent variables: percentage of the surplus allocated to the Advantaged member of the pair (Columns (1) and (3)), and percentage of the surplus that the dictator kept for him/herself (Column (2)). Standard errors clustered at the individual level in parentheses. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. List of controls: age, gender (man, woman, other), political affiliation (5 categories), education (6 categories), income (7 categories), continent (4 categories), attention treatment (3 categories), task type (4 categories), slider orientation (2 categories).

Figure 2.1: Allocation by Status treatment.



The average allocations to the Advantaged member by Status (Advantaged or Disadvantaged), in both Involved decisions (left) and Impartial decisions (right). The error bars represent 95% confidence intervals based on participant-level data aggregated across trials.

Impartial Allocations.

In the Impartial allocation task, we removed the self-interest of the dictators. Any remaining favoritism by the Advantaged dictators toward Advantaged recipients thus measures the persistence of a self-serving bias that self-interest cannot explain. The right part of Figure 2.1 shows evidence for such a bias, as allocation differences persist into the Impartial trials, with the Advantaged dictators still giving significantly more to Advantaged members of the pair ($p < 0.001$, rank-sum test). Column 3 of Table 2.2 shows that Advantaged dictators gave 3.4 percentage points more of the surplus to the Advantaged member after controlling for individual characteristics. While statistically significant, these differences in Impartial allocations are about one-third of the difference in the Involved trials. Konow (2000) attributes these remaining differences in impartial divisions to shifting norms of fairness, an explanation we investigate in more detail below.

Result 1. *Advantaged dictators gave a larger share of the common account to themselves than Disadvantaged dictators gave to Advantaged recipients or themselves. These differences in allocations persist for Impartial choices, although the effect is less than half the size.*

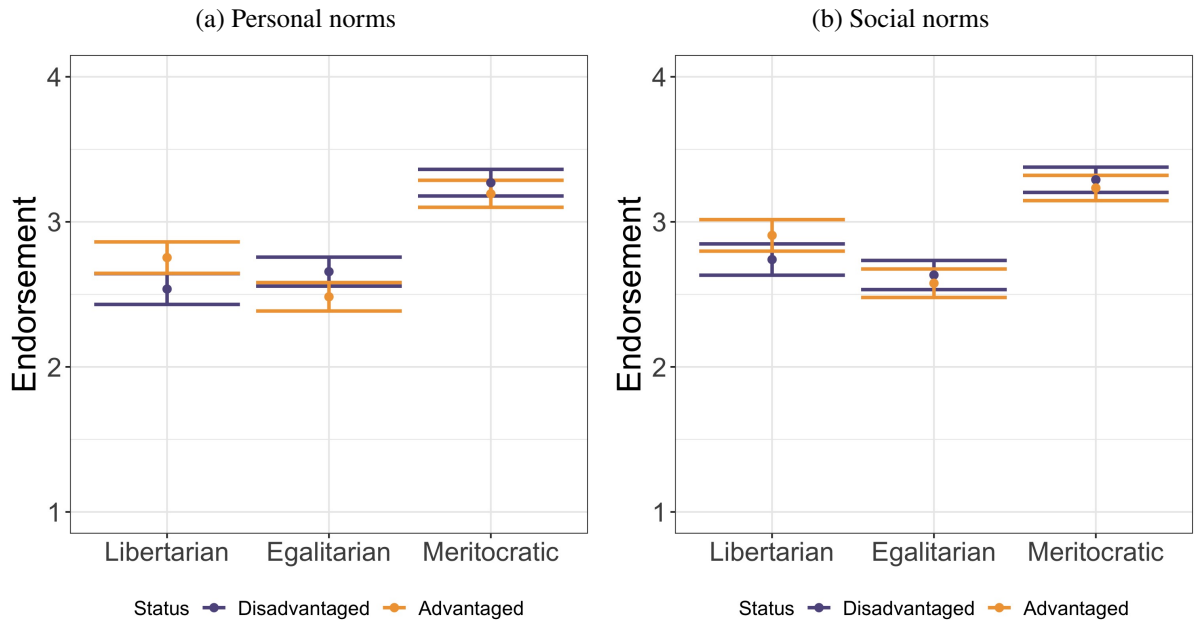
2.4.2 Status, Norms and Beliefs

In this section, we investigate whether dictator Status shifted dictators' beliefs and attitudes related to allocations. In particular, we examine three sets of outcome variables: personal fairness norms, social norms, and beliefs about relative performance.

Personal and Social Norms.

We first look at both personal fairness norms and anticipated social norms about the appropriateness of different fairness criteria. Figure 2.2 shows the effect of Status on norm endorsement, and Table 2.3 gives the results of ordered logit regressions with the discrete appropriateness rating as the dependent variable (see Online Appendix B.1.1 for the effect of Status on our secondary norm elicitations). We find that Advantaged dictators rated libertarian norms as more appropriate on average. A rank-sum test shows the distribution of endorsement is significantly different for both personal norms ($p = 0.0044$) and social norms ($p = 0.026$). The difference in personal norms is confirmed in regressions (Table 2.3, Column 1), but the effect on social norms is smaller and insignificant. In addition, we find that Advantaged dictators were less likely to endorse egalitarian norms, but with statistical significance only for the personal norms elicitation (rank-sum test, $p = 0.015$), a result confirmed in our regression analyses (Table 2.3, Column 1). We find no statistical differences for meritocratic norms. The difference between personal and social norms indicates that subjects had some understanding that their own appropriateness ratings were biased, a finding we explore further below.

Figure 2.2: Personal and social norms split by Status.



The average endorsement of libertarian, egalitarian, and meritocratic norms, both personal (left panel) and social (right panel) split by Advantaged or Disadvantaged Status. Endorsement is measured on a 1-4 scale, with 1 being very morally inappropriate and 4 being very morally appropriate. Libertarian norms mean dividing according to outcomes, including merit and luck. In contrast, egalitarian norms mean splitting evenly regardless of merit or luck. Finally, meritocratic norms mean dividing according to merit alone. Personal norms are those that participants endorse for themselves, whereas social norms are those that they predict others will endorse. The error bars represent 95% confidence intervals.

Table 2.3: Effect of Status on norms

	Personal Norms (1) All data	**Social Norms** (2) All data
Panel A: Libertarian		
Advantaged	0.39*	0.29
	(0.16)	(0.16)
Panel B: Meritocratic		
Advantaged	-0.20	-0.17
	(0.16)	(0.16)
Panel C: Egalitarian		
Advantaged	-0.40*	-0.16
	(0.16)	(0.15)
Observations	600	600

Data from Experiment 1 and Experiment 2. All models are ordered logits. Dependent variable: social or moral acceptability of a norm (1 very inappropriate, 2 somewhat inappropriate, 3 somewhat appropriate, 4 very appropriate). Robust standard errors in parentheses. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. List of controls: age, gender (man, woman, other), political affiliation (5 categories), education (6 categories), income (7 categories), continent (4 categories), attention treatment (3 categories).

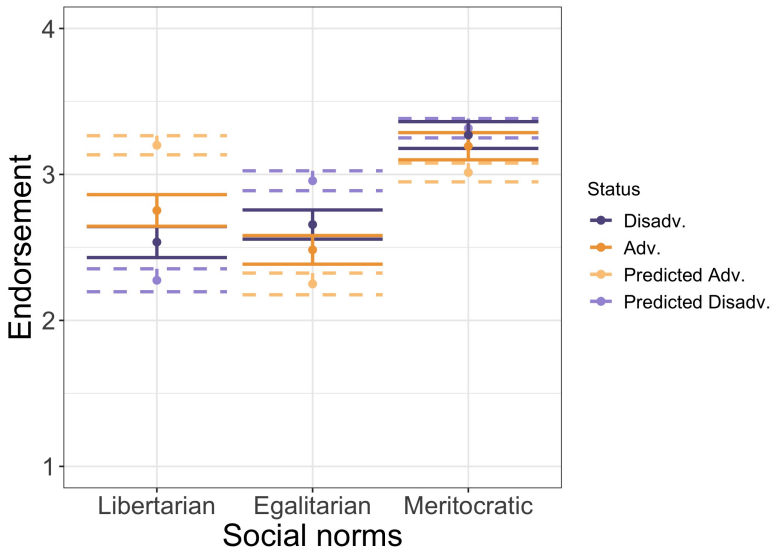
Can participants predict the effect of Status on norms?

We investigate whether participants were aware of the effects of Status on personal norms by asking them to predict social norms separately for Advantaged and Disadvantaged dictators. One hypothesis is that those who fail to see both perspectives and thus do not acknowledge a status bias may show stronger self-serving biases, whereas those who are aware of bias in others may reflect more and exhibit less bias (Babcock and Loewenstein, 1997). Furthermore, an awareness of how Status impacts personal norms might make people more open to interventions that attempt to reduce bias, at least for others.

To test whether participants accurately predicted the status gap in personal norms, we compare personal norms and predicted social norms in Figure 2.3. We find that participants, regardless of Status, correctly anticipated that Advantaged dictators endorsed libertarian norms more highly (rank-sum tests $p < 0.001$) and Disadvantaged endorsed egalitarian norms more highly (rank-sum test $p < 0.001$). This finding suggests that participants know how Status can bias fairness views. In fact, as seen in Figure 2.3, they overestimated the status biases in social norms compared to the actual differences observed in personal norms and further predicted that the Disadvantaged would be more likely to endorse meritocratic norms (rank-sum $p < 0.001$), suggesting that they anticipated others to have stronger status biases than themselves.

Despite the awareness of how Status influenced self-serving biases in allocations, we do not find any relationship between predicting larger status gaps in norms and allocation choices (Spearman's correlations between the predicted gap and the % allocated to the Advantaged in Impartial decisions: libertarian gap: ρ = -0.07, $p = 0.07$; meritocratic gap: ρ = -0.005, $p = 0.91$; egalitarian gap: ρ = 0.02, $p = 0.67$). This lack of relationship between predicted status gaps in

Figure 2.3: Predicted social norms vs. actual personal norms split by Status.



The average endorsement of libertarian, egalitarian, and meritocratic norms. Predicted social norms for Advantaged vs. Disadvantaged and actual personal norms are displayed. Endorsement is measured on a 1-4 scale with 1 being very morally inappropriate and 4 being very morally appropriate. The error bars represent 95% confidence intervals. Note: one's own Status has minimal effect on the predictions of social norms by Status, so the predictions have been collapsed across participants' Status.

social norms and allocations suggests that participants may be subject to similar self-serving biases in allocations that they predict in others. Nevertheless, the awareness that Status impacts fairness norms may matter – despite the lack of correlation with one's own bias – as it could lead to acceptance of interventions to reduce bias in "others", even if people think that they are uniquely immune to such biases.

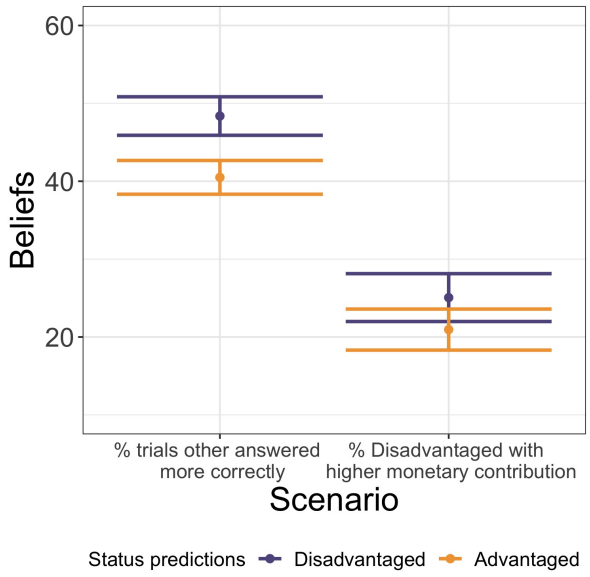
Beliefs about relative performance.

We now turn to beliefs about relative performance, elicited only in Experiment 2. Figure 2.4 shows an overview of the mean beliefs and confidence intervals across Status treatments. All participants showed some bias toward underestimating the number of rounds in which the recipients outperformed them (the true answer was 50% for all participants except 1 for whom it was 40%), but this is particularly pronounced in Advantaged dictators. In line with the tendency to form motivated beliefs $\sim$ 80% of Advantaged dictators indicated that the other participant did equally well or worse than them, compared to only 60% of Disadvantaged dictators (rank-sum test $p < 0.001$). This difference in beliefs is confirmed by an OLS regression displayed in Table 2.4, Column 1. It shows that Advantaged dictators believed that recipients answered more questions correctly in 1.6 fewer rounds (about 8% of the total rounds) than Disadvantaged dictators did ($p < 0.001$). Because beliefs about performance were asked after the allocation decisions, participants could simply have remembered their performance on each round to answer this question, which may reduce the bias in beliefs compared to studies with more ambiguity

(Valero, 2021; Deffains, Espinosa and Thöni, 2016). Nevertheless, the difference in beliefs depending on the Advantaged Status suggests that biased beliefs (or memories) are still present to some extent even with full information, as was also found in Espinosa, Deffains and Thöni (2020).

We then look at beliefs about the size of the disadvantage, as measured bys beliefs about the probability that a Disadvantaged member would out-contribute an Advantaged member. Both groups of subjects overestimated this probability: the real chance was 6.8% while they believed it to be 23% (t-test, $p < 0.001$). In addition, we do not find support for the idea that Advantaged dictators underestimated their random advantage more than Disadvantaged dictators did to downplay the role of luck (rank-sum test $p = 0.068$). An OLS regression analysis (Table 2.4, Column 2) finds that the beliefs go in the opposite direction of what would be expected: Advantaged dictators thought it less likely (by about 4 percentage points) that Disadvantaged dictators outperformed Advantaged dictators in terms of monetary contributions ($p = 0.05$). A plausible alternative explanation is that this result represents a different form of self-serving bias, whereby Advantaged dictators interpreted larger monetary contributions as signifying a higher deservingness instead of a larger artificial advantage.[7]

Figure 2.4: Beliefs about performance by Status.



The average beliefs about performance split by Status (Advantaged or Disadvantaged). On the left, the participants state their beliefs about the % of Involved trials on which the recipients answered more questions correctly than them. In contrast, on the right, the participants estimate the % of trials on which Disadvantaged participants had a higher monetary contribution than Advantaged participants. The error bars represent 95% confidence intervals.

[7]Yet another explanation is that Advantaged dictators were so convinced of being better at the task that their beliefs about the size of advantage did not reverse this. We can check this interpretation in the same model shown in Column 2 by controlling for the dictator's beliefs about the number of rounds in which the Disadvantaged member of the pair answered more questions correctly than the Advantaged member. This additional control does not change the results from Column 2, indicating that the beliefs about the recipients' correct answers are not driving the result.

Table 2.4: Effect of Status on beliefs

	(1) # RecOutperf	(2) % DisOutcont
Advantaged	-1.60***	-4.12*
	(0.35)	(2.09)
Observations	400	400

Data from Experiment 2. All models are linear regressions. Dependent variables: (1) # RecOutperf: the dictators beliefs about the number of rounds in which the recipient answered more questions correctly of the dictator him/herself. (2) % DisOutcont: Beliefs about the % chance that any Disadvantaged dictator contributed a higher monetary contribution than an Advantaged dictator on any round. Clustered standard errors clustered at the individual level in parentheses. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. List of controls: age, gender (man, woman, other), political affiliation (5 categories), education (6 categories), income (7 categories), continent (4 categories), attention treatment (3 categories).

Result 2. *Advantaged dictators personally endorsed egalitarian sharing rules less and libertarian sharing rules more than Disadvantaged dictators. Overall social norm perceptions were similar, but statistically non-significant; however, for social norms split by Dis(Advantage), participants predicted significant status biases regardless of their own Status. Advantaged dictators were also more likely to believe that they outperformed and out-contributed in the task.*

2.4.3 Do Norms and Beliefs Explain Allocations?

We now try to quantify how much of the self-serving bias can be explained by variations in norms and beliefs. As our measure of self-serving bias, we use the effect of Status on allocations in Impartial trials. Since self-interest has been eliminated as a motive in these trials, this status effect is most likely to reflect internalized shifts in fairness or beliefs. We perform this analysis using the common Difference of Coefficients Approach for mediation analysis (Judd and Kenny, 1981).

Correcting for measurement error.

A key problem in mediation analyses comes from measurement error. Gillen, Snowberg and Yariv (2019) show that small noise in the measurement of the mediators – in our case norms and beliefs – can lead to a severe underestimation of their mediating role. To address this problem, we leverage our multiple elicitations in Experiment 2, where we have repeated elicitations of both personal and social norms.

For personal and social norms, we exploit the Instrumental Variables (IV) approach suggested by Gillen, Snowberg and Yariv (2019) to isolate the common variation in the multiple elicitations. More precisely, we instrument our measures of personal norms with the alternative elicitation of the appropriateness of using different types of information to divide the common account, which measure the same underlying construct as discussed in Section 2.3. As an additional instrument, we use our coding of the open-ended answers about the use of fairness criteria. To instrument the social norms, we used the participants' predictions about the social

norms of a) the Advantaged dictators and b) the Disadvantaged dictators. These are valid instruments as they jointly provide information on the perceptions of social norms endorsed in the universe of dictators.

For beliefs, we cannot follow the same approach: our two belief elicitations concern related but not completely overlapping aspects of performance. Hence, the best we can do is to enter both our beliefs measures linearly as controls. As Gillen, Snowberg and Yariv (2019) show, this approach should still reduce concerns due to measurement errors, but it leaves more room for error.

Online Appendix B.1.4 quantifies the importance of correcting for measurement error in our setting by comparing estimates with and without corrections. It shows that we would underestimate the explanatory power of personal norms by a factor of 2.5 without correction for measurement error.

Mediation Results.

Table 2.5 displays the results of linear regressions using the data from Experiment 2, in which we have multiple elicitations of both personal and social norms. This analysis uses observations averaged at the individual level because the variability in norms and beliefs is only between and not within subjects. We first establish our baseline result without controlling for norms or beliefs. Column 1 shows that the overall bias in Experiment 2 is just over 3 percentage points. Note that this is similar to the estimate of 3.4 over both experiments presented in Table 2.2, indicating that the exclusion of Experiment 1 does not change the results.

We next examine whether personal norms, social norms, or beliefs explain the self-serving bias in allocations by Status. Column 2 displays a two stages least square estimation that includes personal norms. In the first stage, the F-statistic is 10.4, indicating that our instruments are highly relevant. To compare the change in coefficient on Advantaged between stages, we use the method for comparing coefficients of nested models described in Clogg, Petkova and Haritou (1995). We find that in the second stage, the coefficient for Advantaged drops to 1.76, is no longer significantly different from zero ($p = 0.090$), and is significantly different from the from 3.04 coefficient in Column (1) ($t(372) = 2.83$), $p = 0.005$). Column 3 displays the same analysis as in Column 2 but with social rather than personal norms. In the first stage, the F-statistic is 21.0, indicating, once again, a small expected bias in the estimates. The coefficient for Advantaged decreases to 2.71, remains both significantly different from zero ($p = 0.001$) and not significantly different from the coefficient in Column 1 ($t(372) = 1.25$, $p = 0.21$). Finally, Column 4 displays a linear regression that controls for beliefs about performance. The coefficient for Advantaged becomes 2.82, remains significantly different from zero ($p = 0.003$ and not significantly different from the coefficient in Column 1 ($t(372) = 0.53$, $p = 0.60$).

We can judge the explanatory power of the three psychological variables by comparing the coefficients for being Advantaged in the different columns. Personal norms explain 42% of the self-serving bias, social norms explain about 11%, and beliefs explain about 6%. To compute these numbers, we use Table 2.5, and we take the difference between the coefficient for

Advantaged in Column (1) and the coefficient for Advantaged in the column where we control for a given psychological channel. We then divide the result for the coefficient for Advantaged in Column (1). For example, the effect of Status that passes via personal norms is given by $(3.04-1.76)/3.04 = 0.42$. Where 3.04 is the total effect of being Advantaged on allocation from Column (1); 1.76 is the effect of Status on the allocation that does not pass via personal norm, from Column (2).[8] The limited explanatory power of social norms is in line with the weak effect of Status on these norms. Instead, the explanatory power of beliefs might be underestimated, as we cannot entirely eliminate measurement bias from the belief variables.

Table 2.5: Impartial allocations to Advantaged recipients controlling for norms and beliefs

	(1)	(2)	(3)	(4)
	% given to Adv.	% given to Adv.	% given to Adv.	% given to Adv.
Advantaged	3.04***	1.76	2.71**	2.82**
	(0.83)	(1.04)	(0.85)	(0.93)
Personal Norms		✓		
		(37.07)		
Social Norms			✓	
			(10.38)	
Beliefs				✓
				(2.27)
F-statistic		10.4	21.0	
Observations	400	400	400	400

Data: Impartial trials from Experiment 2. Dependent variables: percentage of the surplus allocated to the Advantaged member of the pair. Columns (1) and (4) are linear regressions, Columns (2) and (3) are 2SLS models. In Column (2), the instrumented variables are perLib, perEga, and perMer; the instruments are our alternative personal norms elicitations. In Column (3), the instrumented variables are socLib, socMer, and socEga; the instruments are participants' perceptions of the social norms of a) the Advantaged dictators and b) the Disadvantaged dictators. Column (4) contains two beliefs variables. The first is "% DisOutcon". The second is "# DisOutperfom", which indicates the dictators' beliefs about the number of rounds in which the disadvantaged member of the pair answered more questions correctly in the task. This variable is generated from a simple transformation of the variable "# RecOutperf". The variables mentioned in this caption are defined in Table 2.1. List of controls common to all regressions: percentage of the joint number of correct answers due to the Advantaged recipient, age, gender (man, woman, other), political affiliation (5 categories), education (6 categories), income (7 categories), continent (4 categories), attention treatment (2 categories), slider orientation (2 categories). The parentheses under the coefficients for "Advantaged" report the standard errors clustered at the individual level; The parentheses below the check marks for report the $\chi^2(3)$ (for norms) and the $\chi^2(2)$ (for beliefs) statistics for the joint significance of these variables. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. The F-statistics is the Kleibergen-Paap rk Wald F statistic.

Do norms and beliefs predict allocations?

Next to the mediation, we look at the direct connection between norms, beliefs, and allocations. We do so by looking at the coefficients for these variables in the regressions described in Table

[8] An alternative approach to compute the mediation effect is based on the product of the effect of Status on norms and of the effect of norms on allocation. This approach described by VanderWeele and Vansteelandt (2014) yields the same results as the one we have presented.

2.5, the same we used for the mediation analysis above. To reduce the risk of false positives, we use 3 joint statistical tests rather than testing the significance of each of the 8 coefficients separately and therefore do not display the individual coefficients in Table 2.5 (these coefficients are reported in a table included in Online Appendix Table B.1.6). Moreover, we compare the p-values with the Bonferroni adjusted significance level of $\alpha = 0.0167$, obtained by dividing the canonical $\alpha = 0.05$ by 3, the number of tests we are running. The test of the joint significance of personal norms in Column (2) rejects the null hypothesis that none of the three personal norms correlates with allocation decisions ($\chi^2(3) = 37.07$, $p < 0.001$). The test on the joint significance of social norms in Column (3) rejects a similar null hypothesis for social norms ($\chi^2(3) = 10.38$, $p = 0.0156$), showing that social norms correlate with decisions as well. Finally, the test in Column (4) fails to reject the null hypothesis that neither of the two beliefs correlates with behavior ($\chi^2(2) = 2.27$, $p = 0.32$). Online Appendix Table B.1.7 shows evidence that the effect of personal norms on behavior depends on the dictator's Status.

Robustness.

Table 2.5 does not include specifications that combine norms and beliefs to estimate the total amount of self-serving biases that are mediated by these variables. The reason is that instrumenting the personal and the social norms at the same time results in a first-stage F-statistic below 2 and, hence, in a large expected bias in the estimates, likely due to collinearity between some of the instruments. An additional limitation is that the analysis assumes a linear relationship between norms, beliefs, and redistribution. To address these limitations, the regressions discussed in Online Appendix B.1.3 control for beliefs entering them as 4th-degree polynomials and for personal and social norms entering them as dummy variables for each possible norm rating. Moreover, to limit the bias due to measurement error, every set of elicitations available for the norms is included in the regression (Gillen, Snowberg and Yariv, 2019). The results of this alternative mediation analysis are similar to the one from Table 2.5. Moreover, this analysis shows that our ability to explain self-serving biases does not change much if we control for personal norms, social norms and beliefs jointly.

Result 3. *Shifts in personal norms capture 42% of the self-serving bias in impartial decisions; social norms capture 11%, whereas self-serving beliefs about performance capture at least 6%.*

2.5 Discussion

A number of design features of our study have implications for the generalizability and interpretation of our findings, which we discuss in detail below.

First, we always elicit all of our constructs after participants make their allocation choices. This means that participants may want to justify their allocations when answering these constructs. In the literature on the order effects of choice and elicitations, d'Adda, Drouvelis and Nosenzo (2016) find that behavior may shift if norms are elicited first but that norms (especially incentivized social norms) are less likely to change regardless of whether they are elicited before

or after behavior, suggesting that our choice to elicit norms after allocation choices is the cleanest design. However, Rustichini and Villeval (2014) find a more a bi-directional relationship between personal norm elicitations and choice where norms elicited after choice may be used to justify decisions. A recent paper by Charness, Gneezy and Rasocha (2021) on eliciting beliefs also discusses mixed evidence on whether elicitations bias subsequent choices, arguing that more research is needed and that, in light of the ambiguity, the more important metric (choice or belief) should be elicited first. We consider allocation choices as our primary measure and thus we elicited it first. Similarly, we elicit beliefs, personal norms, and social norms in the same order across all participants. We chose the order to try to minimize spillover effects by putting questions that might be perceived as more leading later in the order of elicitation. Nevertheless, there could be some influence of earlier questions on later ones that is not fully accounted for in our analyses.

Second, a few features of our design may push toward personal norms explaining the most variance. One feature is the anonymous setting which eliminates social stakes such as reputation or punishment, giving personal norms the best chance of influencing behavior (Fehr and Gächter, 2000; Eckel, Fatas and Kass, 2022; Salazar et al., 2022). The anonymity or visibility of context, in addition to other situational factors like the salience of personal or social norms, likely affects the extent to which people feel obligated to follow social norms (Kallgren, Reno and Cialdini, 2000; Cialdini, Kallgren and Reno, 1991). The lack of anonymity may be why Bursztyn, González and Yanagizawa-Drott (2020) and Sparkman, Geiger and Weber (2022) find that social norms dominate personal norms in settings where the fear of sanctions may reduce the relative importance of personal norms. Ajzen and Fishbein (1970) also show that even the simple framing of a prisoner's dilemma as competitive or cooperative can impact the relative weight of social norms vs. personal norms. That said, more social exposure could also reduce bias in social norms, as the subjects would have a bigger incentive to correctly anticipate the reactions of others to their choices. Given that we were primarily interested in the justifications people use even when there are no consequences because decisions like voting about redistribution are typically taken in private, we designed the study to focus on self-image in an anonymous setting. Another design feature potentially making personal norms more important is that personal norms are not incentivized, allowing them to be more subject to a consistency bias and justification since there is little cost to manipulating them, as opposed to social norms, which are incentivized.

Finally, the participants make the involved allocations before the impartial ones, which may strengthen the biased allocations in the impartial decisions. We are interested in how developing self-serving fairness principles in the involved decisions spills over into impartial decisions due to cognitive dissonance, hence the choice of the ordering. However, putting the impartial allocation before the involved could show whether cognitive dissonance works in the other direction, whereby participants stick to more impartial fairness rules even in the involved decisions (or shift their fairness rule from the beginning, anticipating the effect on involved decisions). In the literature, Dengler-Roscher et al. (2018) directly test how the order of involved vs. impartial decisions affects allocations after a real-effort task, albeit in a situation without luck. They find

larger deviations from meritocratic divisions for involved allocations as compared with impartial allocations and less deviation from these meritocratic divisions when impartial allocations are made first (but only for participants who do not have prior experience of allocation tasks). Further, Valero (2021) shows that knowing there will be a later opportunity to redistribute doesn't shift beliefs about the underlying luck or merit of success, suggesting that people are not strategic enough to manipulate their beliefs when they could financially benefit from it later. On the contrary, Saccardo and Serra-Garcia (2023) show that the formation of an unbiased opinion reduces subsequent corruptibility of experimental subjects when giving financial advice, but they also find that subjects anticipate such effects. Together, these findings suggest that putting impartial allocations first might reduce the effect of status on norms and self-serving allocation biases, an avenue for further research.

2.6 Conclusion

In this paper, we investigate the role of norms and beliefs in explaining self-serving biases. We find evidence that randomly advantaged participants are less likely to believe that redressing inequalities due to luck is morally appropriate and more likely to overestimate their economic performance. However, the random advantage leads to smaller and insignificant shifts of social norms. Variation in norms and beliefs explains around 42% of the self-serving bias in allocation behavior, primarily driven by the impact of personal norms. Our design allows precise quantitative estimates thanks to a reduction in measurement error, which more than doubles the impact of such norms relative to uncorrected estimates.

These results show that economic status has an effect on personal norms as well as beliefs, with shifts in personal norms of fairness emerging as the most important explanation of self-serving biases. This suggests that modeling efforts should focus on this particular psychological mechanism. So far, there does not seem to be consensus as to how to best incorporate such norms in economic models. Policy-makers who aim to reduce self-serving biases about redistribution should also focus on personal norms, for instance through moral persuasion campaigns that have been successful in reducing ethnicity-based biases (Blouin and Mukand, 2019). While rewiring people's conceptions of what constitutes socially acceptable behavior might be difficult to accomplish and less impactful, our findings suggest that campaigns can be effective by targeting personal norms, which are relatively elastic.

While personal norms can explain a large part of self-serving biases, almost 60% of the bias in our experiment remains unexplained. One additional factor not explored in this study is the potential for in-group favoritism, which has been found to play an important role in differential allocations in prior studies (Dorin et al., 2021; Cassar and Klein, 2019). Future work may further reduce measurement bias in beliefs and account for other mechanisms, like in-group favoritism, that this study does not explore. Moreover, this paper performs a correlational mediation analysis that does not allow for causal claims on the relationship between the mediators – norms and beliefs – and behavior (Imai, Tingley and Yamamoto, 2013). Future work might

study these relationships more directly, by developing experimental designs that manipulate beliefs or norms.

Chapter 3

Memory Sophistication

3.1 Introduction

Memory mistakes can cost a person's savings, career, or life. Investors can forget information that affects the value of a company and misprice it. Referees can misremember the details of old articles and mistakenly reject a manuscript. Doctors can forget the telltale symptoms of a dangerous complication and give a wrong diagnosis. Kahana (2012) presents an overview of the many ways in which human memory fails, while more and more economists are investigating memory limitations and their consequences (see for example Mullainathan, 2002; Zimmermann, 2020; Bordalo et al., 2023).

This paper studies whether people are sophisticated about their memory limitations and asks three questions. Are people overconfident about the accuracy of their current recollections? Can they predict how good their memory will be in the future? And does the complexity of the memory task affect sophistication? The paper calls agents "sophisticated" if they can correctly assess the accuracy of their memory, "overconfident" if they think that their memory is better than it is, and, vice versa, "underconfident" if they think that their memory is worse than it is.

The first question is important because for memory to explain mistakes, people need to be overconfident about the accuracy of their recollections. Only if they are, will they take what they remember at face value when making a decision. If they are not, they will return to their notes or the Internet, or, when cross-checking is too costly, they will reduce their exposure to memory mistakes. Sophisticated investors can reduce the capital they invest; sophisticated doctors can keep a patient overnight for further monitoring. Indeed, several articles assume that people are overconfident about their memory (Mullainathan, 2002; Fudenberg, Lanzani and Strack, 2022). This assumption seems reasonable: we know that people are overconfident about their ability (Camerer and Lovallo, 1999), their self-control problems (Toussaert, 2018), and present bias (DellaVigna and Malmendier, 2004). Yet, this paper shows that memory sophistication is a complicated phenomenon and that people are not always overconfident.

With the second question, I focus on the moment people learn that a piece of information will be helpful in the future. At this moment, people can write notes or rehearse the information to reduce their future memory losses. Sophisticated agents can correctly trade off the cost and

benefit of a memory investment. Overconfident ones are likely to invest too little, increasing the scope for memory mistakes. Underconfident ones might waste time and energy on information that they are already likely to remember.

Finally, the third question investigates the relationship between sophistication and the complexity of the memory task. This relationship is relevant because both the vertical and horizontal organization of a firm change the characteristics of the memory task the employees have to complete - for example, the number of pieces of information that workers have to remember increases with the number of tasks they have to complete. Hence, understanding the effects of complexity on sophistication can help predict which tasks are more likely to generate memory errors, and it can pave the way to optimize the work tasks to minimize the workers' mistakes.

I answer these questions with an experiment that runs over two stages, separated by several days. The first stage presents a list of associations between a prompt and a number, and it asks participants how likely they are to remember these associations during the second stage. The second stage takes place several days after the first. It tests the subjects' memory, and it elicits their confidence in the accuracy of their recollection. The experiment focuses on associative memory because this kind of memory underpins many important decisions. Doctors must associate symptoms with diseases, investors have to connect news with affected companies, and referees need to associate manuscripts with relevant articles. The design does not restrict the memory-enhancing technologies the participants can use - including taking notes and rehearsing the information. As such, the experiment measures people's ability to assess the efficacy of whatever mnemonic strategy they choose.

Three between-subjects treatments shed light on the effects of complexity on memory sophistication. In one treatment, the participants have to remember only 2 associations. In the second one, the associations increase to 7. In a final treatment, the number of associations goes back to 2, but the memorizing is made more difficult by the presence of other but pay-off irrelevant associations - these irrelevant associations increase what the psychologists call "interference". Orthogonally to these treatments, the experiment varies the similarity of the prompts, a third dimension of complexity.

Similarity, interference, and the number of associations to remember are interesting to study because many on-the-job memory tasks differ on these dimensions of complexity. Managers need to remember more facts than their analysts and face more interference because not all their memories are relevant to each decision. Investors have to memorize more similar information when they specialize in a sector of the economy. Moreover, the importance of these dimensions is enhanced by their underpinning economic biases. Bordalo et al. (2023) and Enke, Schwerter and Zimmermann (2022) show how similarity and interference can generate systematic mistakes like over- and under-reaction to information, the availability and representativeness heuristics, the overestimation of the probability of unlikely events, and the dependency of beliefs on the order in which the information is acquired.

The main finding is that memory sophistication is a complicated phenomenon. Some memory tasks produce overconfidence, but others generate underconfidence. A higher complexity of

the memory task can either exacerbate underconfidence or make people overconfident depending on the dimension of complexity that increases.

More in detail, I find that the participants make systematic mistakes in predicting the accuracy of their future memory. When the task is the simplest, the participants are underconfident about their future memory. This underconfidence increases when the participants have to remember more associations, possibly because the participants underestimate the value of practice. Instead, interference reduces underconfidence. Finally, similarity has a U-shaped effect on sophistication. In the treatment where the participants have to remember only two associations and there is no interference, they are closer to sophistication if they face the most similar or dissimilar prompts. Overall, the average deviations from sophistication are sizable as they reach up to 38% of the biggest possible mistake. Yet in all the treatments, the participants are closer to sophistication than to complete under or overconfidence.

The mistakes persist when the participants retrieve past information to make a decision. Participants are underconfident when they have to remember many associations. Instead, they are overconfident when there is interference: they realize that interference makes the task harder, but they underestimate how much harder the task became. Once again, the errors are considerable: the average deviation from sophistication reaches 23% of the biggest possible mistake. In addition, I find that, with time, people become more confident in all treatments. This upward shift in confidence mitigates ex-ante underconfidence but exacerbates ex-ante overconfidence.

Overall, the results indicate that we can expect memory mistakes to be economically meaningful when interference makes memorizing harder. Moreover, they suggest that people might waste too much time improving their memory, especially when they have to remember many associations - a mistake I document for the first time. As I discuss in the next Section, these results contribute to the economic literature on memory, sophistication, and complexity. In Section 3.6 I discuss the implications of the findings for the workplace, for the way students are tested, and for social learning.

3.2 Literature Contribution

This paper adds to several strands of literature. First, it links to the growing work that focuses on the role of memory in economic decisions. This literature has studied how memory limitations can underpin biases in judgment and valuations (Mullainathan, 2002; Bernheim and Thomadsen, 2005; Bordalo, Gennaioli and Shleifer, 2020; Bordalo et al., 2023; Enke, Schwerter and Zimmermann, 2022; Fudenberg, Lanzani and Strack, 2022) and how limited memory is used to develop motivated beliefs (Bénabou and Tirole, 2002; Zimmermann, 2020; Saucet and Villeval, 2019; Huffman, Raymond and Shvets, 2022; Chew, Huang and Zhao, 2020; Gödker, Jiao and Smeets, 2021; Müller, 2022).

Some papers assume the agents to be sophisticated about their memory recollections (Bernheim and Thomadsen, 2005). Others that people take their memories at face value (Mullainathan,

2002). A third group explicitly analyses the relationship between overconfidence and the consequences of limited memory: in Fudenberg, Lanzani and Strack (2022) overconfidence can exacerbate long-run memory biases, while in Bénabou and Tirole (2002) the level of overconfidence changes the number of equilibria. Finally, some papers study the relationship between memory and beliefs when people cannot verify the accuracy of their memories (Bordalo et al., 2023; Enke, Schwerter and Zimmermann, 2022).

The first contribution of this paper is to show that people are closer to sophistication than to complete overconfidence. Yet, overconfidence is large when people have to deal with interference. This last remark suggests that the finding that interference generates several costly mistakes by Bordalo et al. (2023) and Enke, Schwerter and Zimmermann (2022) extends to settings where people can cross-check their recollections.

Second, the paper contributes to the well-established empirical literature on agents' sophistication. This literature has focused on sophistication about present bias (Augenblick and Rabin, 2019; Le Yaouanq and Schwardmann, 2022), about self-control problems (DellaVigna and Malmendier, 2006; Toussaert, 2018), in strategic interactions (Nagel, 1995; Crawford and Iriberri, 2007), and about cognitive abilities different from memory (Camerer and Lovallo, 1999; Schwardmann and Van der Weele, 2019). Still, it has so far mostly neglected sophistication about memory limitations.

In economics, two papers provide evidence on memory sophistication. In the final experiment of Enke, Schwerter and Zimmermann (2022), the participants must decide how much to bet on some companies. To correctly remember the company's value, they need to overcome interference and similarity. In this setting, the participants can reduce the amount they bet if they are unsure about their memory. Yet, they don't seem to exploit this risk management option. On the contrary, the subjects that suffer from more biased memory bet larger amounts. These results suggest that people are overconfident about their memory limitations.

My paper confirms that people can be overconfident when there is interference but also shows how they are underconfident if interference is not there. Memory underconfidence is a novel and important finding: it suggests that people might waste too much time verifying their memories. Moreover, and differently from Enke, Schwerter and Zimmermann (2022), this paper does not restrict the mnemonic strategy people can use. Hence, it shows that under and overconfidence can emerge even when the participants are free to choose how to remember the information.

Bronchetti et al. (2022) also provides some evidence on memory sophistication. It finds that people are overconfident about the probability that they will complete an action within the required time window. Instead, I study sophistication about the probability of accurately remembering a piece of information, which is a relevant dimension of sophistication for understanding doctors', investors', and referees' mistakes. Moreover, I identify factors that favor or depress confidence.

Third, the paper adds to a nascent literature on the relationship between complexity and cognitive mistakes (Li, 2017; Kendall and Oprea, 2021; Oprea, 2022; Enke and Graeber, 2019). Here the contribution is to show how interference, similarity, and the number of associations,

all factors known to make memorizing harder, affect the accuracy of people's beliefs about the quality of their memory.

Fourth, the paper makes a small contribution to the literature on motivated reasoning. This literature has shown that people want to hold positive views about themselves and it has examined imperfect memory as a mechanism people can leverage to develop rosy beliefs (see Bénabou and Tirole, 2016, for a review). This paper suggests that memory ability is in itself an ego-relevant trait.

Finally, in creating a connection between the literature on sophistication and the one on memory, this paper relates to the psychological work on metamemory (see Dunlosky and Tauber, 2016, for a review).

3.3 Design

The experiment took place over two stages, several days apart. During the first stage, the participants learn some associations; during the second, they are tested on their memory of this information and receive a bonus if they remember it correctly.

The experiment allows me to manipulate one dimension of complexity at a time and to retain control over the incentives to remember. These two features are essential for answering the research questions. Yet, they are impossible to achieve in the field settings that motivate this study. Across professions, the memory tasks simultaneously vary in difficulty and in the incentives for accurate recollections.

The experimental instructions are in Appendix C.2. The experiment was preregistered. The preregistration document and a discussion of a deviation from it are in Appendix B.1.

3.3.1 Stage 1

In Stage 1, the participants are presented with color-number pairs organized in a list. The color acts as a cue. The participants have to memorize the number associated with each of them. The colors come from a colorblind-friendly palette, while the numbers are two-digits integers. After seeing the list, the participants face a practice task that forces them to memorize the associations. They see the colors of the list one by one in random order, and they need to type the number associated with each of them. They receive feedback, and they repeat the practice till they manage to twice correctly associate every color of a list with its number. This practice phase ensures that inattention is not a driver of the results.

Following this practice phase, the participants read that they will complete a Recall Task during the second stage of the experiment and that they could win a bonus if they completed this task correctly. At this point, the participants cannot access the information about the list anymore, and they have to rely on their memory to complete the Recall Task. In this way, the experiment is similar to the many instances where people acquire information before learning the exact value of remembering it. For example, an academic might come up with a new idea and only then realize the importance of a conversation he had months before. In addition, the

impossibility of going back to the source of information makes the experiment similar to the many life instances - like informal meetings, listening to the radio while driving, or social gatherings - where information can be acquired, but it is hard to save the information anywhere else than in one's memory.

The Recall task involves typing the number associated with each color in the list. The computer then randomly selects one of the colors, and the participants will win the bonus if their answer for this one color is correct. The bonus will be either of £2 or £3 with equal probabilities, with the exact value revealed only at the end of the second stage of the experiment.

Next, I offer the participants the opportunity to buy a computer code that assures them that they will win the bonus in the second stage even if they don't remember the color-number pairs correctly. The computer code checks the participants' answers to the Recall Task and corrects them if they are wrong. As such, the computer code acts as a memory-enhancing tool. However, the participants learn whether they have gotten the compute code only after they have completed the Recall Task. Hence, they all need to indicate their recollections.

I elicit the participants' willingness to pay (WTP) separately for the case in which the bonus is equal to £2 and the case in which the bonus is £3. To do so, I use two separate Multiple Price Lists (MPL), one per each possible bonus level. The participants know that there is a 10% chance that the WTP is pay-off relevant and that with the remaining probability, they will not receive the computer code independently of their answers. The WTP provides a link between beliefs and behavior.

After the WTP elicitation, I ask the participants to assess their probability of winning the bonus during the second stage of the experiment conditional on them not receiving the computer code. Given the incentives structure for accurate memory, this question is equivalent to asking which share of the associations they think they will remember correctly during Stage 2. The elicitation happens after the two MPLs, allowing the participants to adjust their predictions depending on their WTP. In fact, the participants' WTP might change the optimal investment in real-life memory technologies: the participants with a higher WTP are likely to spend less time and effort rehearsing the color-number combinations and hence have less accurate recollections during Stage 2.

Since the participants can use any mnemonic strategy they like, their beliefs measure their confidence in the effectiveness of the strategy they thought optimal. This design feature increases the external validity of the results. Outside the experiment, people can often increase the probability of remembering with external aids - for example, they can write a note. Yet these mnemonic strategies can fail - the note might be lost. Hence, the memory sophistication that matters in real life includes the sophistication about the effectiveness of external memory aids.

The belief elicitation is not incentivized to avoid introducing distortion in participants' answers and behavior, as any incentive scheme that links rewards to the accuracy of the guess changes the incentive to have an accurate memory.[1] In not incentivizing the beliefs, this pa-

[1]The use of incentivized belief elicitations as commitment devices is investigated in Augenblick and Rabin (2019), which finds mixed evidence about it. Toussaert (2018), develops an incentivized method that leverages a participant's beliefs about another similar subject. However, her method does not fit well with the research

per joins a growing list of papers that use the same strategy to measure self-control (Ameriks et al., 2007), economic preferences (Falk et al., 2018; Enke, Rodriguez-Padilla and Zimmermann, 2022), behavioral biases (Stango and Zinman, 2020), and cognitive uncertainty (Enke and Graeber, 2019). Another reason not to incentivize the predictions is that the participants should have already thought about these beliefs when they expressed their WTP for the computer code: the probability of remembering correctly is a key determinant of the value of the code. Section 3.5.4 shows that, despite the lack of incentives, the beliefs data in the experiment are of high quality.

The first stage of the experiment concludes with a survey. The survey asks the participants about their demographics, their memory abilities, and their mnemonic strategies in real life. It also asks them to explain how they decided upon their WTP for the computer code.

3.3.2 Stage 2

In the second stage, the participants complete the Recall Task. Afterward, I ask them to assess the probability that each of their answers in the memory task is correct. These beliefs are also not incentivized lest the additional instructions about incentives interfere with memory processes.[2]

The experiment concludes by providing the participants with feedback about their performance in the Recall Task and their earnings.

3.3.3 Manipulations of complexity

The experiment manipulates complexity with two orthogonal between-subjects manipulations. The first one varies the number of associations and the amount of interference. The second changes the information similarity.

Baseline memory task. The simplest treatment is called "*2 Pairs*". In it, the participants have to memorize one list of two color-number pairs.

Increasing the number of associations. In the "*7 Pairs*" treatment, the number of pairs grows to seven. This increase makes the memory task more complex because people can recall a smaller percentage of associations when they need to remember a larger number of them (Kahana, 2012).

Introducing interference. In a third treatment called "*2 Pairs + Interference*", the participants have to learn two lists. The first one is made of two pairs. The second is made of seven.

question at hand, as it is not clear why an overconfident subject should predict that another subject will be equally overconfident.

[2] Adding these instructions before the Recall task risked introducing a common shock in participants' memory ability, possibly due to unexpected fatigue. Adding them between the Recall Task and the belief elicitation inserts a time gap incompatible with measuring beliefs when a judgment is formed.

Only the first list is relevant for the Recall Task, while the second list aims to distract the participants and make remembering harder. This treatment increases complexity by exploiting what psychologists call “interference": the forgetting generated by memorizing new but irrelevant associations.[3]

The treatment displays the list sequentially. The participants see the second list only after they have completed the practice of the first. Only after completing the practice of the second list, the participants are informed about the Recall Task and that the task will test only their memory of the first list.

Changing the similarity of the cues. The design manipulates between subjects the similarity of the cues - that is, of the colors - in the associations. These manipulations happen in the *2 Pairs* and *2 Pairs + Interference* treatments, where the participants have to remember only two color-number pairs.

To vary the color similarities, for each subject, the experiment randomly selects two colors from a palette made of seven. The procedure generates 21 different combinations. The distance between each pair of colors is calculated using the CIEDE2000, the industry standard to asses color similarity (Luo, Cui and Rigg, 2001). The calculation results in distances between 19.8 and 83.5 out of a theoretical range of 0 to 100. All the values are above the threshold that indicates that people can easily distinguish the two colors (Mokrzycki and Tatol, 2011).

Similarity makes remembering harder because it introduces a form of interference even when participants have to remember a single list. During Stage 2, a color might cue the recollection of both numbers making it hard for the participant to choose which of the two is the relevant one. The probability of both numbers coming to mind should increase with the similarity of the prompts (Kahana, 2012). This effect of similarity on memory is one of the building blocks of the model by Bordalo et al. (2023).

3.3.4 Implementation

Sample and data collection. The final sample comprises 957 subjects that completed both days of the experiment. 1218 subjects started the first day of the experiment, 999 completed it and were invited to the second session.[4]. The participants were recruited on Prolific.co, an online platform, from 22nd to 29th of September, and the two experimental days were 5 days apart.[5]. Of the participants that concluded the experiment, 331 are in the *Easy* treatment, 312

[3]The colors of the first list also appear in the second one, further increasing the interference the participants face and generating a versions of the A/B A/C memory association task commonly used in psychology (see Kahana, 2012) and implemented also by Enke, Schwerter and Zimmermann (2022).

[4]An additional 667 started the first day of the experiment, but I could not invite them to the second day because of a server failure that made their data unreliable. 3 additional subjects completed the experiment but then returned their submission, possibly because they didn’t want to complete the second session anymore.

[5]For 81 participants the two sessions were 6 days a part. These were the only subjects recruited on 22/09 who were not affected by the server failure. This difference in the break length does not pose any identification threat: the participants always completed Stage 2 one the date that was communicated to them at the beginning of Stage 1.

are in the *Many Items* treatment, and 315 are in the *Interference* treatment. The experiment run on custom-made PHP codes.

50% of the participants are females. The average age is 40 years old. I accepted only participants based in the UK that did not report any issue with color visualization. They earned a completion fee of £3.75 plus an average bonus of £1.71, and they took less than 29 minutes to complete Stage 1 and 5 minutes to complete Stage 2.

Implementation details. To ensure that the experimental subjects understood all the instructions' essential elements, I used slides that displayed the instructions step by step. The slides describe in detail the Recall Task (including a screenshot), the incentives behind it, the probability of winning the bonus depending on the number of correct answers, the computer code and its characteristics, and the probability that the WTP elicitation is pay-off relevant. Many slides included an explanatory image to make the instructions easier to digest.

20 comprehension questions tested the participants' understanding of these features. I did not allow the subjects to continue with the experiment until they answered all the questions of each set correctly. These extensive quizzes ensure that any lack of sophistication is not due to misunderstandings.

During Stage 2, I sent two reminders to complete the experiment, and I offered online assistance to participants with any problem. These two measures helped to keep attrition below 5%.

3.4 Results

Summary statistics. Table 3.1 summarizes the means and the quartiles of the most important outcome variables. I analyze these data in detail below, but some high-level observations are noteworthy. First, memory is not perfect in the experiment. On average, the participants recall 63% of the pairs correctly. This figure reassures us that the Recall Task was not trivial and that many participants did not have access to infallible memory-enhancing technology outside the experiment. The participants' mistakes are unsystematic. The average deviation from the correct answer is insignificantly different from zero as the mistakes of people recalling numbers that are too high are canceled by the mistakes of people recalling numbers that are too low ($p = 0.10$).[6]

Second, the difficulty of the Recall Task changes with the treatments. The *2 Pairs + Interference* treatment is the most challenging. There, the average recall rate is 41%, and less than 35% of the participants remember all the associations. Unexpectedly, the *7 Pairs* treatment exhibits the highest average recall rate with 77% of successful recalls. Yet only 42% of the participants in this treatment have perfect memory. In comparison, the *2 Pairs* treatment has a lower average recall rate (71%), but it is the easiest treatment in the sense that 67% of the subjects recall all the associations correctly.

[6]This analysis excludes the 1.9% of guesses below 10, the lowest possible correct answer. These low guesses most likely come from participants who are sure they don't remember the correct number and type only one digit to quickly progress with the experiment.

Table 3.1: Summary Statistics

		Mean	Q1	Median	Q3	% Underconfident	% Overconfident
Share correct recalls							
	2P	0.71	0.00	1.00	1.00		
	7P	0.77	0.71	0.86	1.00		
	2P+I	0.41	0.00	0.00	1.00		
Ex Ante Beliefs							
	2P	0.58	0.50	0.50	0.87		
	7P	0.48	0.14	0.50	0.80		
	2P+I	0.46	0.20	0.50	0.51		
Ex Ante Sophistication							
	2P	-0.13	-0.50	-0.10	0.10	54%	29%
	7P	-0.30	-0.60	-0.23	0.00	75%	19%
	2P+I	0.05	-0.30	0.05	0.50	34%	52%
Ex Post Beliefs							
	2P	0.71	0.50	0.80	1.00		
	7P	0.70	0.56	0.75	0.90		
	2P+I	0.54	0.25	0.50	0.87		
Ex Post Sophistication							
	2P	0.00	-0.18	0.00	-0.02	44%	26%
	7P	-0.08	-0.18	-0.04	0.00	58%	29%
	2P+I	0.13	0.00	0.05	0.32	25%	57%

"2P" indicates the *2 Pairs* treatment which has 331 complete observations. "7P" indicates the *7 pairs* treatments which has 311 complete observations. "2P+I" indicates the *2 Pairs + Interference treatment* which has 315 complete observations.

The high recall rate in the *7 Pairs* treatment might be due to the design of the learning phase of the experiment during Stage 1. This phase concludes only when the participants can reconstruct *all* the color-number pairs. The high number of pairs in this treatment makes it hard to recall all of them and forces participants to repeat the practice task multiple times (5.3 times on average VS 2.2 in the other treatments). This repetition, in turn, might make it easier to remember at least *some* of the pairs. Indeed the lowest quartile in this treatment has a recall rate of 71% VS the 0% of the same quartile in the other treatments.

Similarity affects forgetting as well. Column 1 of Table C.1.1 in the Appendix regresses the fraction of correct answers in the *2 Pairs* treatment on the distance between the colors and the square of this distance. The first-order term is positive, indicating that the memory task is easier when the colors are further away from each other ($p = 0.043$). At the same time, the second-order term is negative ($p = 0.071$), generating an inverse U-shaped relationship between similarity and recall rate. Column 2 of the same table indicates that no similar relationship appears in the *2 Pairs + Interference* treatment, perhaps because the effect of interference overrides the one of similarity.

Selective attrition. The experimental design is vulnerable to selective attrition. When it introduces interference or increases the number of pairs to remember, it forces the participants to put more effort into the practice phase of Stage 1. Moreover, across treatments and similarity levels, the experiment varies the difficulty of the Recall Task, creating differential incentives to

complete Stage 2 of the study.[7]

In the *7 pairs* and the *2 Pairs + Interference* treatments, where the required effort is highest, the participants are indeed more likely to drop out. Of the 1093 participants that reached the point where the treatment was assigned, 8% gave up before the end of Stage 1.[8] The attrition rate is significantly higher in the *7 pairs* treatment (11%) and in the *2 Pairs + Interference* treatments (9%) than in the *2 Pairs* treatment (4%) ($p < 0.001$ and $p = 0.017$, Fisher's exact test). However, there is no significant difference in attrition between the *7 pairs* treatment and in the *2 Pairs + Interference* treatments ($p = 0.33$ Fisher's exact test).[9] In addition, the variations in the similarity of the colors do not cause different attrition rates within the *2 Pairs* and the *2 Pairs + Interference* treatments as shown in columns 3 and 4 of Appendix Table C.1.1.

No additional selective attrition occurs before or during Stage 2. Only 4.5% percent of the participants that completed Stage 1 failed to complete Stage 2, all of them because they did not start the second survey despite the reminders. The percentage oscillates between 4% and 5% with no significant difference across treatments ($p = 0.93$, in a Fisher's exact test that compares all three treatments).[10] In this stage as well, similarity does not affect completion rates (see columns 5 and 6 of Appendix Table C.1.1).

To sum up, there is no evidence of differential attrition between the *7 Pairs* treatment and the *2 Pairs + Interference* treatments, so these two treatments can be compared with a simple t-test. However, we observe lower dropout rates in the *2 Pairs* treatment due to the simplicity of the memory task in this condition. This selective attrition does not endanger the preregistered analysis: this analysis compares the empirical level of sophistication with the theoretical benchmark of complete sophistication in each treatment. Furthermore, Section 3.4.4 use conservative assumptions about the participants that drop out to investigate the effect of the different dimensions of complexity of sophistication.

3.4.1 Ex Ante Sophistication

To study whether the participants can correctly predict the accuracy of their future memory, I create a variable called "Ex Ante Sophistication". Ex Ante Sophistication is the difference between a participant's Stage 1 beliefs about her probability of winning the bonus in the Recall

[7]This vulnerability is inevitable given the research questions. As discussed above, the practice phase is needed to ensure that inattention is not a driver of the results. Moreover, the study needs to be between subjects. As such, any increase in complexity comes with a drop in the incentives to complete the study. There was no existing evidence to gauge the size of the attrition elasticity with respect to complexity in advance.

[8]Almost all the dropouts happened while participants learned the associations or during the comprehension questions about the Recall Task and computer code.

[9]Stage 1 completion times are similar in the two treatments (19 seconds difference, $p = 0.97$) suggesting that the effort needed to complete these two treatments is comparable.

[10]This low attrition rate might be due to the steep incentives the participants faced. With less than 5 minutes of work, they earned a minimum of £3.75 (the combined completion fee for the entire experiment), and they avoided their submission to the first session was rejected, something that would reduce their probability of being recruited for future studies on Prolific.

Table 3.2: Sophistication and beliefs shifts

	(1)	(2)	(3)	(4)
Panel A: Ex Ante Sophistication				
	All data Ex Ante Soph.	2P Ex Ante Soph.	7P Ex Ante Soph.	2P+I Ex Ante Soph.
Average	-0.12***	-0.13***	-0.30***	0.050*
	(0.015)	(0.027)	(0.022)	(0.025)
Observations	957	331	311	315
Panel B: Ex Post Sophistication				
	All data Ex Post Soph.	2P Ex Post Soph.	7P Ex Post Soph.	2P+I Ex Post Soph.
Average	0.019^{+}	-0.0016	-0.077***	0.13***
	(0.010)	(0.018)	(0.012)	(0.020)
Observations	957	662	2177	630
Panel C: Beliefs updating from day 1 to day 2				
	All data Δ Beliefs	2P Δ Beliefs	7P Δ Beliefs	2P+I Δ Beliefs
Average	0.14***	0.13***	0.22***	0.085***
	(0.012)	(0.020)	(0.021)	(0.019)
Observations	957	331	311	315

All models are linear regressions with a constant as the unique variable. In panels A, C, and in Column 1 of panel B, a participant is an observation; in all the other columns of panel B, a color-number pair is an observation. Column 1 uses all the observations, Column 2 uses the ones from the *2 Pairs* treatment, Column 3 the ones from the *7 Pairs* treatment, Column 4, the ones from the *2 Pairs + Interference* treatment. Dependent variable: in Panel A, Ex Ante Sophistication; in Panel B, Ex Post Sophistication; in Panel C, absolute beliefs change between Session 1 and Session 2. Robust standard errors in parentheses in Panels A, C, and in Column 1 of Panel B; Clustered standard errors at the individual level in Columns 2, 3, and 4 of Panel B. $^{+}$: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

task - a variable I call “Ex Ante Beliefs" - and her share of correct recalls during Stage 2. I.e.,

$$\text{Ex Ante Sophistication} := \text{Ex Ante Beliefs} - \text{Share correct recalls}.$$

The difference between these two variables captures the accuracy of the participants' predictions because the probability of winning the bonus is 1 if the participant remembers all the associations, it is 0 if she doesn't remember any, and it is linearly increasing in the number of correct recalls between these two extremes. Ex Ante Sophistication is equal to zero if a participant can perfectly predict her future memory perfectly. It is positive if they are overconfident and negative if they are underconfident.

Across all treatments, the average Ex Ante Sophistication is equal to -0.12, indicating that the participants underestimate by 12 percentage points the share of associations that they can recall correctly; an equivalent interpretation is that the participants underestimate the probability of winning the bonus by 12 percentage points. Panel A of Table 3.2 regresses Ex Ante Sophistication on a constant, and, in Column (1), it shows how this average level of underconfidence is statistically different from zero ($p < 0.001$).

There is, however, substantial heterogeneity in Ex Ante Sophistication across treatments.

Panel A of Table 3.2 indicates that treatments *2 Pairs* and *7 Pairs* exhibit significant underconfidence. Participants in the *2 Pairs* treatment underestimates the share of associations that they are able to recall correctly by 13 percentage points ($p < 0.001$) - they predict to remember 58.1% of the pairs when they actually remember 71.0%. The underestimation grows to 30 percentage points for the participants in the *7 Pairs* treatment ($p < 0.001$) - 47.7% predicted VS 77.5% actual recall rate. On the contrary, the participants in the *2 Pairs + Interference* treatment overestimate the share of correct recalls by 5 percentage points ($p = 0.049$) - 46.4% predicted VS 41.4% actual recall rate. Table 3.1 corroborates these findings showing that the majority of subjects is too pessimistic in the *2 Pairs* and *7 Pairs* treatments, but it is too optimistic in the *2 Pairs + Interference* treatment.

When we compare the *7 Pairs* and the *2 Pairs + Interference* treatment, the two treatments with similar attrition rates, we see that Ex Ante Sophistication is significantly higher in the latter ($p < 0.001$). We can then conclude that the characteristics of the memory task influence memory sophistication. Furthermore, this finding indicates that the relationship between sophistication and complexity is not straightforward. Some dimensions of complexity promote underconfidence, others overconfidence. Section 3.4.4 delves deeper into this result.

To benchmark the findings above, the left panel of Figure 3.1 displays the average deviation from sophistication as a percentage of the biggest possible mistakes.[11] It shows that the deviations are between 8% and 38% of their largest possible value. While these deviations are of a considerable magnitude, they also reveal that people know their memory is neither perfect nor helpless. The smallest deviation happens in the *2 Pairs + Interference*, while the biggest is in the *7 Pairs* treatment. In the *2 Pairs* treatment, the average deviation is equal to 18% of the largest possible mistake.

Result 1. *People anticipate that their future recollections will not be perfect. Yet, depending on the characteristics of the memory task, they are either under or overconfident about their future memory.*

3.4.2 Ex Post Sophistication

Next, I analyze whether people have a good sense of how accurate their memory is when they have to recall past information. To this end, I define the variable “Ex Post Sophistication" as the difference between the Stage 2 beliefs that a given recall is correct - a variable I call “Ex Post Beliefs" - and an indicator variable equal to 1 if the recall is correct.

$$\text{Ex Post Sophistication} := \text{Ex Post Beliefs} - \mathbb{1}_{\text{correct recall}}.$$

[11] To find the largest possible mistake, I observed that within a treatment, the highest possible value of the average Ex Ante Sophistication equals 1 minus the average share of correct recalls. This characterizes a population that erroneously believes to be able to remember all the associations with certainty. Vice versa, the lowest possible value of the average Ex Ante Sophistication is equal to -1 times the average share of correct recalls, and it characterizes a population who is erroneously sure that it will not remember any association.

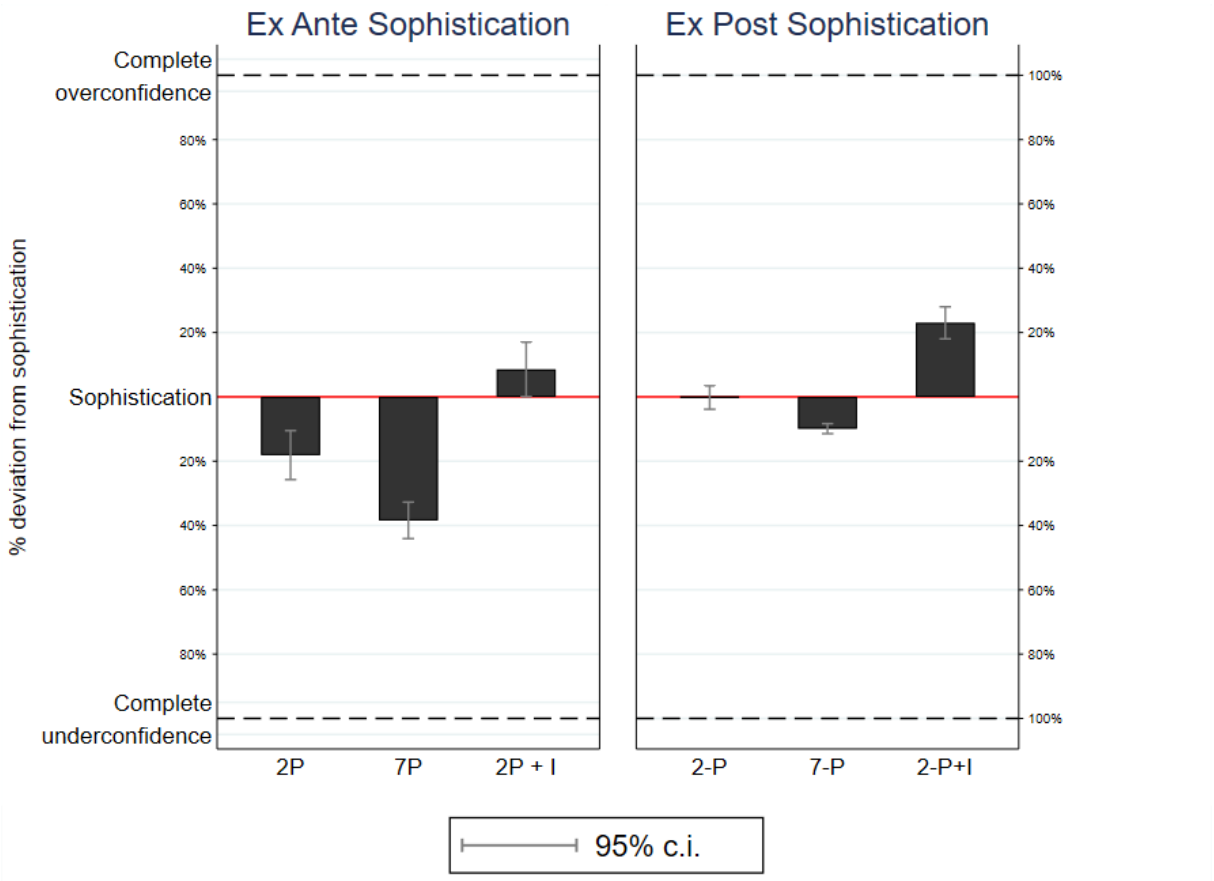


Figure 3.1: The left panel shows the average deviations from Ex-Ante Sophistication; the right one the average deviations from Ex-Post Sophistication. The data is normalized such the average deviation would have been equal to 1 if every subject had erroneously believed to have perfect memory (Complete overconfidence) and equal to -1 if every subject had erroneously believed to have zero chances to remember the color-number pairs correctly (Complete underconfidence). This normalization consents to asses how close to complete under or overconfidence people are on average. In both panels: the left bar uses the data from the *2 Pairs* treatment, the central bar uses the data from the *7 Pairs* treatment, and the right bar uses the data from the *2 Pairs + Interference* treatment. Height of the bar: average deviation from sophistication as a percentage of the maximum possible deviations. Bars: 95% confidence intervals.

In the *2 Pairs* and *2 Pairs + Interference* treatments I obtain 2 observations of Ex Post Sophistication per subject, one observation for each association. In the *7 Pairs* treatment, I obtain 7 observations per participant.

A positive average Ex Post Sophistication indicates that participants are overconfident about the precision of their recalls. Vice versa, a negative average Ex Post Sophistication signifies underconfidence.

Taking the average individual Ex Post Sophistication and pooling the data from all treatments, people overestimate by 2 percentage points their share of correct answers.[12]. Panel B of Table 3.2 regresses Ex Post Sophistication on a constant and, in Column (1), it indicates that this slight deviation from sophistication is only significant at the 10% level ($p = 0.065$).

Panel B of Table 3.2 shows significant heterogeneity in Ex Post Sophistication across treatments. In the *2 Pairs*, treatment the participants' predictions are spot on. On average, they underestimate by only 0.2 percentage points. This is a precisely estimated insignificant deviation from perfect sophistication: with 95% confidence, the average mistake is smaller than 3.7 percentage points in either direction. In the *7 Pairs* treatment, the participants are underconfident on average. They think they remember 70% of the associations correctly when they actually remember 77% of them. This 7 percentage points underestimation is highly statistically significant ($p < 0.001$), and it realizes as 58% of the participants are too pessimistic about their performance. Finally, in the *2 Pairs + Interference* treatment, participants overestimate their fraction of correct recalls by 13 percentage points ($p < 0.001$) - 54% predicted VS 41% actual recall rate. This overestimation is driven by the 57% of participants that are too overconfident.

There is a significant difference in Ex Post Sophistication between the *7 Pairs* and the *2 Pairs + Interference* treatment ($p < 0.001$) which implies that, depending on the characteristics of the memory task, people can be either under or overconfident about the accuracy of their current recollection their share of correct answers.

The right panel of Figure 3.1 displays that the average deviation from sophistication is equal to 10% of the deviation characterizing a completely naive population in the *7 Pairs + Interference* treatment and to 23% of the same benchmark in the *2 Pairs + Interference* treatment. Similarly to the results for Ex Ante Sophistication, these deviations are of considerable magnitude, but they reveal that people are closer to perfect sophistication than to complete naivete.[13]

Result 2. *When people form judgments, they realize they don't fully recollect past information. People have an accurate perception of the accuracy of their memory when they have to remember few associations. However, people can be under or overconfident for more complicated tasks depending on the task characteristics.*

[12] Averaging at the individual level is necessary not to overweight the participants from the *7 Pairs* treatment for which I obtain 7 rather than 2 data points for Ex Post Sophistication.

[13] See Footnote 11 for an explanation on how to calculate the value of the largest possible deviations.

3.4.3 Belief shift between Stage 1 and Stage 2

Between Stages 1 and 2 of the experiment, beliefs about the share of correct recall shift upwards. To display this result, Panel C of Table 3.2 uses the variable "Δ Beliefs": the difference between the Ex Post Beliefs averaged at the individual level and the Ex Ante Beliefs. I. e.,

$$\Delta \text{ Beliefs} := \text{subject's average Ex Post Beliefs} - \text{Ex Ante Beliefs}.$$

Pooling all the treatments, beliefs shift upward by 14 percentage points on average ($p < 0.001$), with 57% of the participants becoming more optimistic and only 30% becoming more pessimistic. This shift in beliefs occurs in every treatment. The upward movement is between 8.5 and 22 percentage points (or between 18% and 45% of the Ex Ante Beliefs, and between 17% and 46% of the variance of the sum of the two beliefs) depending on the treatment, and it is always significant the 0.1% level. As a result of the upward shift in beliefs, the participants in the *2 Pairs* and *7 Pairs* treatment, who were on average underconfident in Stage 1, get closer to sophistication in Stage 2. Vice versa, the participants in the *2 Pairs + Interference* treatment, who were already overconfident during Stage 1, move further away from sophistication.

There are two possible explanations behind the upward shift in beliefs. The first is learning. Most participants are underconfident during Stage 1. With time they might get a better sense of the quality of their memory and update their beliefs upward. To provide evidence of this mechanism, Column 1 of Appendix Table C.1.9 regresses the beliefs shift on Ex Ante Beliefs and the share of correct recalls. The coefficient for the share of correct answers is positive and highly significant ($p < 0.001$). Since the regression includes the Ex Ante Beliefs as a control, this result supports the learning story: conditional on the initial beliefs, the participants that remembered more observations shifted their beliefs to a larger extent. Column 2 in the same table adds demographic controls to the regression and confirms the result.

The second possible explanation is motivated reasoning. The participants might like to think that their memory is good and forget negative signals about it - a selective forgetting that makes them more optimistic (Zimmermann, 2020). A piece of evidence supporting this channel is that the upward shift worsens the overconfidence in the *2P+I* treatment, a finding inconsistent with learning. Another piece of evidence comes from Appendix Table C.1.9. On top of Ex Ante Beliefs and the share of correct recalls, the regressions in this table include a dummy equal to one for the participants who consider memory more important in their work and daily activities than the median participant. These participants prize their memory and might derive a higher utility from believing that it is good. If motivated reasoning contributes to the belief shift, we should expect these participants to exhibit a larger belief jump. Indeed, in Column 1, the coefficient for this dummy indicates that these participants shift their beliefs upwards by 3.2 percentage points more than the others ($p = 0.028$). In Column 2 which includes the demographic controls, the coefficient drops slightly but remains significant at the 10% level ($p = 0.061$).

Overall, both learning and motivated reasoning contribute to the upward shift of beliefs between Stage 1 and Stage 2 of the experiment.

Result 3. *Over time people become more optimistic about the accuracy of their memory. This upward shift brings people closer to sophistication if they are ex-ante underconfident. But if they are ex-ante overconfident, it pushes them away from sophistication.*

3.4.4 Complexity and memory sophistication

In this section, I analyze how similarity, interference, and the number of associations to remember affect people's memory sophistication.

Number of associations. To study the effect of increasing the number of associations on sophistication, I compare the *7 Pairs* to the *2 Pairs* treatment. As discussed in Section 3.4, the comparison is confounded by treatment-specific selection. To address this issue, I use conservative assumptions to impute the sophistication of the participants that dropped out during Stage 1. In the *2 Pairs* treatment, I assume that the dropouts are as underconfident as the 10th percentile of the distribution of sophistication in that treatment. Vice versa in the *7 Pairs* treatment, I assume that the dropouts are as overconfident as the 90th percentile of the distribution of sophistication in that treatment.

These assumptions are conservative, and they go against finding any treatment effect. In the *2 Pairs* treatment, all but one dropouts happen after the participants have already learned the color-number pairs. These dropouts are likely due to the comprehension questions about the willingness to pay elicitation. Hence, there is no reason to expect these participants to be more or less sophisticated than average. Moreover, I assume that the dropouts in the *7 Pairs* treatment are among the most overconfident subjects in their treatment. Yet, these participants likely stopped the experiment because they struggled to remember all the associations correctly during the practice phase. Their giving up suggests that they believed they had little chance to remember the associations correctly in a reasonable amount of time, and it makes them unlikely to have been very overconfident about their memory.

Table 3.3 reports the results of the treatment comparisons under the assumption discussed above. Column (1) in Panel A indicates that the *7 Pairs* treatment reduces Ex Ante Sophistication by 8 percentage points, exacerbating the underconfidence already present in the *2 Pairs* treatment ($p = 0.019$). This effect realizes because the participants in the *7 Pairs* treatment predict to remember less, but they remember more pairs correctly than the ones in *2 Pairs*, as shown in Table 3.1. Their mistake might come from a lack of appreciation for the effect of practice during Stage 1. As discussed at the beginning of this section, the participants spend more time practicing the association in the *7 Pairs* treatment. This interpretation resonates with the underestimation of the benefit of learning from experience documented by Billeter, Kalra and Loewenstein (2011) and Le Yaouanq and Schwardmann (2022).

Column (1) in Panel B indicates that the treatment difference becomes insignificant by Stage 2, suggesting that time mitigates the underestimation of the benefits of practice ($p = 0.17$).

Table 3.3: Complexity and Sophistication

	(1)	(2)	(3)	(4)
Panel A: Ex Ante Sophistication				
	Data: 2P and 7P Ex Ante Soph.	Data: 2P and 2P+I Ex Ante Soph.	Data: 2P Ex Ante Soph.	Data: 2P+I Ex Ante Soph.
7 Pairs	-0.082*			
	(0.035)			
2 Pairs + Interference		0.097**		
		(0.037)		
Color distance			-0.024**	-0.0026
			(0.0083)	(0.0084)
Color distance^2			0.00024**	0.000011
			(0.000083)	(0.000089)
Panel B: Ex Post Sophistication				
	Data: 2P and 7P Ex Ante Soph.	Data: 2P and 2P+I Ex Ante Soph.	Data: 2P Ex Ante Soph.	Data: 2P+I Ex Ante Soph.
7 Pairs	-0.029			
	(0.021)			
2 Pairs + Interference		0.072**		
		(0.027)		
Color distance			-0.0076	0.0053
			(0.0059)	(0.0063)
Color distance^2			0.000068	-0.000052
			(0.000061)	(0.000064)
Observations	700	695	331	315

All models are linear regressions. Column 1 uses the observations from the *2 Pairs* and *7 Pairs* treatments. Column 2 uses the ones from the *2 Pairs* treatment and *2 Pairs + Interference* treatments. Column 3 uses the data from the *2 Pairs* treatment, while Column 4 uses the ones from the *2 Pairs + Interference* treatment. In Columns 1 and 2, for the participants that dropped out in the middle of session 1, I have imputed a sophistication value equal to either the 10th or the 90th percentile of their treatment, depending on which of the two percentiles would make the treatment effect *smaller*. This means that in Column 1, I have imputed the value of the 10th percentile from the *2 Pairs* treatment to the missing observations of that treatment, while I have imputed the value of the 90th percentile from the *7 Pairs* treatment to the missing observations of that treatment. Instead that in Column 2, I have imputed the value of the 90th percentile from the *2 Pairs* treatment to the missing observations of that treatment, while I have imputed the value of the 10th percentile from the *2 Pairs + Interference* treatment to the missing observations of that treatment. Color distance is computed using the CIEDE2000 formula. Robust standard errors are in parentheses. $^{+}$: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

However, this null result must be interpreted cautiously due to conservative assumptions behind the analysis. Indeed, under slightly less conservative assumptions, the worsening of underconfidence persists during Stage 2. For example, we would have found that underconfidence is 5 percentage points ($p = 0.011$) more severe in the *7 Pairs* than in the *2 Pairs* treatment if we had assumed that the dropouts in the *2 Pairs* treatment were as sophisticated as the average participant in their treatment.

Result 4. *When the participants have to remember more associations, they are further away from Ex Ante Sophistication as their underconfidence increases. Instead, the effect of the number of associations on Ex Post Sophistication is less clear.*

Interference. Here, I compare the *2 Pairs* and the *2 Pairs + Interference* treatments to identify the changes in sophistication due to an increase in interference. This comparison is polluted by treatment-specific selection as well. To address the confound, I use assumptions similar to the ones above. In particular, I assume that the dropouts in the *2 Pairs* treatment are as overconfident as the 90th percentile of the distribution of sophistication in that treatment. Vice versa for the *2 Pairs + Interference* treatment, I assume that the dropouts are as underconfident as the 10th percentile of the distribution of sophistication in that treatment.

These assumptions are conservative. I now assume that the dropouts in the *2 Pairs* treatment are overconfident. In contrast, I assumed that these same participants were underconfident when studying the relationship between sophistication and the number of associations. These two assumptions contradict each other, but I maintain them to make the analysis more conservative. Besides, I assume that the dropouts in the *2 Pairs + Interference* treatment are among the most underconfident in their treatment. Yet, underconfidence is possible only if people perform well in the task, and it is unlikely that the dropouts *2 Pairs + Interference* would have performed well had they completed the study. Most of them gave up while trying to learn the interference pairs, suggesting they needed more effort than average to do so. As the interference pairs are meant to be a distraction, the participants that struggled with them should perform poorly in the Recall Task.[14]

Under the assumptions above, Column (2) in Panel A of Table 3.3 shows that interference increases Ex Ante Sophistication by almost 10 percentage points ($p = 0.008$). This happens as the participants realize that interference makes the Recall task more difficult, but they underestimate the extent of the increased difficulty, as shown in Table 3.1. After imputing the values for the missing observations, Ex Ante Sophistication is still significantly smaller than zero in the *2 Pairs* treatment ($p < 0.001$), but it becomes indistinguishable from zero in the *2 Pairs + Interference* treatment. Hence, I can conclude that interference reduces ex-ante underconfidence but not that it generates ex-ante overconfidence.

[14]For the dropouts to be as ex-ante underconfident as I assume they are, they would have needed, for example, to remember 50% of the pairs in Stage 2 while predicting to not be able to remember any. Such a high recall rate is implausible: the participants in the *2 Pairs + Interference* who completed the experiment recalled only 41% of the pairs.

Column (4) indicates that the treatment difference persists during Stage 2 ($p = 0.008$), where interference increases overconfidence by 7 percentage points. In this case, there is enough evidence to conclude that interference generates overconfidence. After imputing the values for the dropouts, average Ex Post Sophistication remains insignificantly different from zero in the *2 Pairs* treatment and significantly positive in the *2 Pairs + Interference* treatment ($p = 0.23$ and $p < 0.001$).

Result 5. *Increasing interference reduces underconfidence and can generate overconfidence.*

Similarity. Finally, I look at the effects of similarity on sophistication. I do so separately for the *2 Pairs* and the *2 Pairs + Interference* treatments. Hence, I don't have to correct for selective attrition.

Column 3 in Panel A of Table 3.3 regresses Ex-Ante Sophistication in the *2 Pairs* treatment on the distance and the square of the distance between the two colors. It finds that Ex-Ante Sophistication decreases with the distance between the two prompts but increases with the square of this distance. Both effects are statistically significant at the 1% level ($p = 0.004$ in both cases), and their combination produces a U-shaped relationship between similarity and Ex-Ante Sophistication. The 20% of participants facing the most similar colors and the 20% facing the most dissimilar one are closer to sophistication, so close that we can't reject the hypothesis that they are perfectly sophisticated. On the contrary, the other 60% of the participants are significantly underconfident. Appendix Figure C.1.1 depicts this U-shape. The U-shape realizes as the combination of two effects. As discussed at the beginning of Section 3.4, in the *2 Pairs* treatment, the task is the hardest for intermediate similarity values. Yet, for these same values, the participants think that the task is the easiest.[15]

The effects of similarity on sophistication disappear in Stage 2. Column 3 in Panel B of Table 3.3 regresses Ex-Post Sophistication in the *2 Pairs* treatment on the distance and the square of the distance between the two colors. It finds that the influence of similarity on Ex-Post sophistication drops to 30% of the one on Ex-Ante Sophistication. The effect is not significant ($p = 0.20$ and $p = 0.26$ for the first and second-order coefficients). The disappearance comes together with a change in participants' beliefs. During Stage 2, the misperception that intermediate levels of similarity make the memory task easier vanishes.[16]

There is no detectable relationship between similarity and sophistication in the *2 Pairs + Interference* treatment. Column 4 in panel A of Table 3.3 regresses Ex-Ante Sophistication on the distance of the two colors and the square of this distance. It shows that the coefficients for these relationships are insignificant ($p = 0.75$ and $p = 0.90$) and one order of magnitude smaller than in the *2 Pairs* treatment. Column 4 in panel B of the same table repeats the analysis for Ex Post Sophistication. It finds that the coefficients are insignificant ($p = 0.40$ and $p = 0.41$). The

[15]This remark emerges from the analysis of the participants' Ex Ante Beliefs. Column 1 of Appendix Table C.1.8 regresses those beliefs on the color distance and the square of this distance. It finds evidence for a U-shaped relationship between similarity and Ex Ante Beliefs.

[16]This result derives from the analysis of the participants' average Ex Post Beliefs. The analysis is displayed in Column 3 of Appendix Table C.1.8, which regresses the Ex Post Beliefs on the color distance and the square of this distance.

lack of a relationship between similarity and interference in the *2 Pairs + Interference* treatment may be unsurprising. As we discussed at the beginning of Section 3.4, there is no significant relationship between similarity and memory accuracy in this treatment.

Result 6. *Absent interference, similarity has a U-shaped effect on Ex Ante Sophistication. The participants are underconfident about their future memory for intermediate levels of similarity but are close to sophistication for extreme levels. This relationship vanishes when the participants judge the accuracy of their current recollections. There is no significant relationship between similarity and sophistication when interference is present.*

3.5 Discussion

This section shows that beliefs about one's memory matter for behavior and that the behavior in the experiment correlates with real-life decisions. Moreover, it presents suggestive evidence about misspecified mental models about the value of memory-enhancing tools, and it shows that individuals' observable characteristics are poor predictors of sophistication. Finally, it discusses and excludes two threats to the study's internal validity.

3.5.1 Beliefs matter for behavior

The paper focuses on people's beliefs about the accuracy of their memory and shows that these beliefs are miscalibrated. But do people use these beliefs to make decisions? This section shows that beliefs about memory limitations indeed matter for behavior.

The first piece of evidence in this direction comes from the questionnaire. 72% of the participants said that the probability of remembering the associations was one of the determinants of their WTP for the computer code. This finding indicates that most participants use their beliefs to decide how much to invest in memory aids.

The second comes from the relationship between Ex Ante Beliefs and WTP. Column (1) in Appendix Table C.1.7 regresses the WTP on people's Ex Ante Beliefs. It shows that the WTP is lower for the participants who predict their memory to be more accurate ($p < 0.001$).[17] This correlation between beliefs and willingness to pay is consistent with participants' using their probability of remembering to evaluate the computer code.

3.5.2 Behavior in the experiment correlates with real-life behavior.

This section presents evidence suggesting that the paper's findings might expand beyond its abstract experimental setting. At the end of Stage 1, the participants completed a survey about

[17] A concern about this relationship is that the direction of causality runs from the WTP to the beliefs. A higher WTP could depress the beliefs because it increases the probability that the participant will get the computer code and, hence, reduces the incentives to have an accurate memory. However, even if the direction of causality moved from WTP to beliefs, the negative relationship between the two variables would support that the beliefs shape behavior. If the WTP changes the Ex Ante Beliefs, it is because it increases the (unmeasured) beliefs that the participant will receive the reward for the task.

their memory abilities and memory strategy in the real world. This section checks whether the participants' behavior in the experiment correlates with their answers to these questions.

Table C.1.5 shows that the participants that report higher memory ability do better in the memory task by 16 percentage points ($p < 0.001$). They also have higher beliefs about their memory accuracy both ex-ante and ex-post ($p < 0.001$ for both comparisons).[18] In addition, the participants that completely agree with the statement "I have a good memory" have, on average, a WTP for the computer code 24 pence lower than the other subjects ($p = 0.036$).[19] Finally, the participants that completely agree with the statement "I usually take notes during work meetings and/or in class" have an average WTP for the computer code 24 pence higher than the other subjects ($p < 0.001$). This last finding shows that the behavior in the experiment correlates with behavior at school and in the workplace.

Yet, not all the evidence points in the same direction. The same Table C.1.5 indicates that the participants who make a list before going grocery don't have a significantly different WTP from the other subjects. The same is true for those subjects that think they are more likely to forget information without the help of lists and notes.

All in all, the evidence suggests that the experimental task relates to behavior outside the lab. Yet, future memory research could look for tasks even closer to real-life behavior.

3.5.3 Additional results

This section presents additional results. The first paragraph presents evidence suggesting that people have wrong mental models about the value of memory-enhancing tools. The second one discusses how individual observable characteristics are poor predictors of sophistication.

Mental models about the value of reminders The results presented so far show that people are not fully sophisticated about their memory limitations; this section indicates that people are likely to make additional mistakes when deciding how much to invest in reminders and other memory-enhancing technologies. It does so by looking at the self-reported determinants of the Willingness To Pay (WTP) for the computer code. Table C.1.2 in the Appendix summarizes the factors the participants considered when deciding on their WTP. 28% of subjects didn't think about the probability of winning the bonus, while 15% disregarded the bonus size available in the Recall Task. In total, 41% of the subjects indicated that they didn't think of at least one between the level of the incentives and the probability that the computer code will be helpful.

The neglect of incentives and the probability of remembering is a strong predictor of WTP. Averaging the two WTP elicitations, Column (1) in Table C.1.3 indicates that the people who think about the probability of remembering the numbers correctly pay 33 pence less for the computer code, while the people that think about the size of the reward pay 71 pence more ($p < 0.001$ for both coefficients). Column (2) in the same table shows the participants that think about the incentives react more to an increase in the bonus payment ($p < 0.001$).

[18] Participants report their memory ability during the first session of the experiment. Hence, their performance in the task cannot influence their answers.

[19] This finding does not appear in Table C.1.5.

Other strong predictors for the WTP are disliking the computer code because it offers unwanted help and liking the computer code because it reduces cognitive load. The first reduces the average WTP by 38 pence ($p < 0.001$). The second increases it by 46 pence ($p < 0.001$). These two effects confirm that the participants who liked the computer code more were willing to pay more for it, reassuring us about the data quality.

Jointly, the four questions about the determinants of WTP explain 22% of the variance in average WTP. This is an impressive fraction considering that these variables are binary and that the complete set of demographic controls can explain less than 2% of the variance.

On top of disregarding important variables, the participants might also struggle to correctly react to changes in the incentives. We can see this struggle focusing on the 33% of participants who said that their WTP was not influenced by disliking the computer code because it offered unwanted help and by liking the computer code because it reduced cognitive load. For these participants, the WTP in the case the bonus is $£3$ should be 1.5 times the WTP in the case the bonus is $£2$. Yet, the ratio between the two WTP is only 1.30, significantly smaller than 1.5 ($p < 0.001$). The ratio increases only to 1.33 and remains significantly smaller than 1.5 ($p = 0.003$) when we look only at the participants that said that they thought about the incentives.

These findings suggest that many subjects have misspecified mental models about the value of memory-enhancing technologies. They disregard either the incentives to remember or the probability that the reminder will be helpful. Moreover, even the ones that think about the incentives underreact when the incentives change. The mental models strongly predict subjects' willingness to invest in these technologies, providing correlational evidence that incorrect models are consequential for people's decisions. Future research should provide more direct evidence of this phenomenon.

Individual predictors of sophistication. This section analyzes which individual characteristics predict sophistication in the experiment. Appendix Table C.1.6 regresses Ex Ante and Ex Post sophistication on demographic characteristics. Column (1) in Panel A shows that older people tend to have lower Ex Ante Sophistication than younger ones. The effect realizes as older participants are significantly better in the task - the share of correct recalls increases by 3.4 percent every 10 years ($p = 0.001$) - but their beliefs are not different from the ones of younger subjects ($p = 0.74$). Instead, gender is not significantly correlated with Ex Ante Sophistication ($p = 0.34$). Column (2) shows that controlling for other demographic characteristics does not alter the results. Given that people are on average underconfident, these results indicate this excessive pessimism is particularly spread among older people.

Column (1) in Panel B indicates that Ex Post Sophistication is 4 percentage points lower among women ($p = 0.049$). This happens because women perform slightly better than men in the Recall Task but are a bit more pessimistic than men about their performance; none of these two differences is statistically significant, though ($p = 0.59$, and $p = 0.20$). Column (2) shows that controlling for other demographic characteristics does not alter the results.

Finally, Table C.1.6 indicates that all the demographic characteristics combined explain less than 3% of the between-subjects variance of both Ex Ante and Ex Post Sophistication. This

result indicates that managers cannot use easily observable characteristics to reduce memory mistakes via targeted reminders or customized work tasks.

3.5.4 Excluding threats to internal validity

The possibility of a common shock. The identification of Ex Ante Sophistication in the experiment relies on the assumption that there is no common shock to the participants' memory between Stage 1 and Stage 2. Unexpected events might happen in the participants' lives, leading to prediction errors. However, these events only threaten identification if they induce correlated mistakes across participants.

The experiment involved UK subjects and ran in a turbulent period for UK politics. On 23rd September 2022, Liz Truss's government announced a mini-budget, a reform plan that sparked high volatility in the financial markets, was retracted in a matter of weeks, and ultimately led to the fall of the government. Did these events generate a correlated shock that confounds the results of this study?

Two observations reassure us that the identification holds. First, 92% of our sample started the experiment on 24/09, after the mini-budget was announced, and completed it on 29/09, before the government took the reform back. Even if the announcement shocked these subjects' memory ability, this shock happened before the beginning of the experiment and persisted throughout. Table C.1.4 in the Appendix shows that all the results about Ex Ante Sophistication replicate in this sub-sample. Second, the 76 participants who started the experiment before the budget was announced are underconfident ($p = 0.006$), and their underconfidence is similar to the rest of the sample (Ex Ante Sophistication of -0.15 VS -0.12, $p = 0.58$), suggesting that the surprising announcement did not significantly affect people's memory.

Beliefs data are of high quality even if the elicitations are not incentivized When experimental economists elicit their participants' beliefs, they often reward accurate answers. They do so to incentivize the participants to think carefully about the variables of interest. In this experiment, I did not incentivize beliefs for the reasons discussed in Section 3.3. Here, I provide evidence that, despite the lack of incentives, the beliefs data are of high quality.

First, I exploit that the participants are indirectly incentivized to develop accurate beliefs by the WTP elicitations for the computer code. The value of the computer code depends on the probability of remembering the associations, and 72% of the participants indicated that they thought about the probability of remembering when stating their WTP. Hence, the data quality of these participants' beliefs should be as good as if directly incentivized.

Panel B of Appendix Table C.1.4 replicates the analysis of Ex Ante Sophistication presented in Panel A Table 3.2 including only this 72% of the participants. The idea is to check whether the paper's results are confirmed in this restricted sample where the beliefs data should be of higher quality. The table confirms that the participants are Ex Ante underconfident on average, and both in the *2P* and in the *7P* treatments ($p < 0.001$ for all three tests). Yet it doesn't replicate the results of the *2P + I* treatment. The point estimate for this treatment indicates that

the participants are overconfident, but the result is not significant ($p < 0.55$). All in all, almost all the results are confirmed in this high-quality sample.

Second, I exploit that the WTP gives an alternative and indirect measure of people Ex Ante Beliefs for a subset of the participants. Appendix Section C.1.1 explains how this measure can be recovered, how it is consistent with the study's main results, but also its limitations.

As a final check, Appendix Table C.1.7 regresses the average WTP for the computer code and the percentage of correct answers in the Recall Task on people Ex Ante and Ex Post Beliefs. It finds that the participants that are more confident have a lower WTP and perform better in the task. These associations are significant at the 0.1% level and indicate that the beliefs are significant predictors of incentivized behavior, further reassuring about data quality.

Overall, the evidence suggests that beliefs are of high quality and that the lack of incentives is not a driver of the results.

3.6 Conclusions and Implications

This paper opens the door to the study of memory sophistication. It shows that it is a complex phenomenon as people can be both under- and overconfident about their memory, with the complexity of the memory tasks driving the direction of the deviations from sophistication. It also provides evidence that beliefs about memory matter for people's investments in memory aids and, hence, that wrong beliefs are consequential.

These findings are important to understand the relationship between memory limitations and mistakes. Before many important decisions, people have the opportunity to revise information at little cost, or they can reduce their exposure to memory errors. Hence, we can expect memory to induce costly mistakes only when a task induces overconfidence. This paper indicates that, while not every task produces overconfidence, interference promotes it.

At the same time, this paper is the first to document that some tasks make people largely underconfident about their future memory, a mistake that becomes more severe when people have to remember many associations. These results suggest that sometimes people spend excessive time and effort to improve the accuracy of their memory, possibly because they under-appreciate the benefit of rehearsal. The results raise the question of whether exams like the university admissions and the bar exam produce severe inefficiencies. These exams require the students to remember large sets of information, and underconfident students might spend too much time and effort rehearsing information they are already likely to remember.

Moreover, the dimensions of complexity studied in this paper find correspondence in job tasks. Both the vertical and horizontal organization of a firm change the amount of information workers must remember and the amount of similarity and interference they face. As such, the results from this paper are a first step to predicting which kind of memory-induced mistakes workers are likely to make depending on the task at hand. Moreover, the finding that these two dimensions shift sophistication in opposite directions suggests that it is possible to design tasks in which the mistakes cancel out and workers are close to sophistication. The paper provides

some evidence that the behavior in its abstract task extends to the workplace, but this evidence is based on the participant's self-assessment. More work is needed to test the ecological validity of these findings and translate them to actual job tasks.

Finally, the paper's results might be relevant for social learning and conflict resolution via communication.[20] They indicate that people's confidence in their recollections depends on the complexity of the environment in which they have learned the information. Hence, they suggest the learning environment affects how open people are to persuasion. In addition, the results suggest that memory is an ego-relevant trait and that people tend to forget negative signals about the accuracy of their memory, as they do with negative signals about their IQ (Zimmermann, 2020). This selective forgetting hints at a new reason why social learning can fail. People like to think that their recollections are accurate even when they are not, and, as such, might undervalue the information shared by others. Future research could formally test these intuitions.

The paper leaves many important questions open. First, other dimensions affect the complexity of a memory task, including the length of the time gap between receiving and using the information. Future research can look at the relationship between these dimensions and sophistication. The relationship between information recency and sophistication seems particularly important as it could provide information about the speed and persistence of the upward shift in confidence observed in this paper. Second, recent theoretical and empirical work has highlighted how people manage or fail to learn about their traits and abilities (Le Yaouanq and Schwardmann, 2022; Ba, 2022; Heidhues, Koszegi and Strack, 2023), future studies should investigate whether, and under which conditions, people learn to be sophisticated about their memory. Finally, this paper provides tentative evidence that people make mistakes when computing the value of memory aids because they don't think about some of the variables relevant to the valuation. Future research could provide more conclusive evidence on this error which is related to the experimental literature studying misspecified mental models (Kendall and Oprea, 2021; Esponda, Vespa and Yuksel, 2020; Kendall and Charles, 2022; Barron and Fries, 2022).

[20]Prominent work studying this topic includes Enke and Zimmermann (2019); Conlon et al. (2022); Babcock and Loewenstein (1997); Schwardmann, Tripodi and Van der Weele (2022)

Chapter 4

Correcting Consumer Misperceptions about CO_2 Emissions

This chapter is based on Imai et al. (2022)

4.1 Introduction

Reducing the emission of greenhouse gases is one of the most pressing challenges of our time. Unfortunately, some potent remedies like carbon pricing, are politically contentious. Instead, policy makers frequently stress the role of information about CO_2 emissions to consumers and producers. For instance, the European Commission's "Farm to Fork Strategy," proposes an extensive carbon labeling strategy, while its "New Consumer Agenda" argues for "more reliable information on sustainability" (European Commission, 2020). In the US, the proposed, but ill-fated, American Clean Energy and Security Act of 2009 contained provisions to study and implement carbon information aimed at consumers (Waxman and Markey, 2009), while the Department of Agriculture and regulatory agencies like the EPA implement greenhouse gas labels for cars, beef, and other products. There is also a corporate interest in carbon labeling, as evidenced by carbon-labeling initiatives from several large European retail chains like TESCO, Casino, and E.Leclerc (Taufique et al., 2022).

The key premise behind information policies is that consumers are motivated to mitigate the climate impacts of their consumption, but might underestimate such impacts. The optimal targeting of information then requires us to identify products and activities for which CO_2 emissions are underestimated by a lot and by those consumers that are highly motivated to invest in mitigation. In this paper, we first combine elicited beliefs about CO_2 emissions and preferences for mitigation in a structural model to identify productive targets for information. We then conduct a field experiment to test whether well-targeted information can lead to more climate-friendly consumption by changing beliefs.

Our first study features a representative survey of US consumers ($N = 1,022$). We use incentive-compatible elicitation techniques to measure both point beliefs and belief distributions about the climate impact of a number of products and actions. We then measure valuations

of carbon emissions for the same consumers, using a willingness to pay for different amounts of carbon offsets, thus producing a "willingness to mitigate" function. We find that consumers generally underestimate carbon impact, with the magnitude of underestimation varying both across people and across product categories. The largest underestimates exist for high-carbon-impact food categories such as beef and coffee. Valuations of carbon emission reductions are positive and relatively high, but marginal valuations decline strongly, leading to a concave willingness to mitigate function.

To make predictions about the impact of correcting consumer beliefs, we use a structural model in which we combine each individual's subjective belief distributions with their elicited willingness to mitigate. We compare this expected willingness to mitigate with a counterfactual where subjective belief distributions have collapsed to the true beliefs about carbon emissions, as measured by the latest scientific estimates. The resulting statistic describes the effect of an information campaign aimed at a particular product as the dollar-tax equivalent of correcting beliefs. For example, informing our participants of the carbon impact of 100 grams of dark chocolate is equivalent to raising the price of chocolate by 4.5 US dollars.

Our model has advantages over using only misperceptions to target information campaigns. First, the model controls for a possible mismatch between who is optimistic about the carbon impact of a given product and who cares about mitigating carbon emissions. For example, if only people with a low willingness to mitigate were optimistic about the carbon footprint of flying, then an information campaign aimed at flying would be impotent. Conversely, targeting groups with a high willingness to mitigate is ineffective if these groups are already well-informed. Second, our model explicitly accounts for the interaction of the shape of the subjective belief distributions and the willingness to mitigate function. In particular, information is more effective when it shifts beliefs along the steeper part of the willingness to mitigate curve. Moreover, concavity of the willingness to mitigate function implies that, holding average beliefs fixed, merely making beliefs more precise should increase mitigation efforts.

In our second study, we conduct an online experiment to test the impact of information provision on the demand for beef and for poultry. While these products are both part of a general food category (meat), beef has almost 10 times the carbon impact of poultry in CO_2 equivalents, mostly due to cow methane production and deforestation associated with the production of cattle feed. Our participants understand that beef is more polluting than poultry, but they think that the difference between them is much smaller than it actually is. In line with this, our structural model, applied to the representative survey data, predicts that while information on beef should have a large impact on demand, the impact on demand by providing information on chicken should be small or non-existent.

We recruited $N = 2,081$ subjects via an online platform and elicited their willingness to pay for a package of meat from a premium online butcher using an incentive-compatible procedure that realized some purchasing decisions and had us send meat to selected participants. In four between-subject treatments, we varied the type of meat (beef vs. poultry) and whether we provide information about the carbon emissions associated with the product in question. All conditions feature prominent mentions of the climate change impact of *some* products in order

to keep the salience of climate change constant across settings and thereby isolate the effect of information on consumption that works through beliefs.

Our intervention is successful in changing consumer perceptions related to the two products. However, we find no evidence that information is effective in changing the demand for either beef or chicken. This null result is true for all subgroups in our sample, and robust among those whose beliefs responded to the intervention. Moreover, we can rule out several explanations for the surprising null effect of information on beef consumption. It is not driven by the information making participants pessimistic about substitute products, by people not consuming much meat being the only ones with optimistic priors, by an overly noisy measure of demand, or by a non-replicable statistical fluke. We also rule out behavioral channels like an intention-action gap and are left to conclude that individuals' meat consumption, unlike more abstract elicitations of green preferences like our willingness to mitigate, simply is not subject to concerns about CO_2 emissions in our participants.

Our results suggest that the current enthusiasm about information provision should be tempered, as shifting beliefs may by itself not be effective in increasing voluntary mitigation. We contribute to two literatures that we will review in turn. The first literature measures misperceptions about the CO_2 emissions associated with consumption and about climate change more generally. Our results imply that the presence of misperceptions alone does not imply the tempting conclusion that correcting misperceptions leads to behavioral change. A second literature documents small, but discernible, effects of carbon labels on consumption. Our results imply that, contrary to the explanation commonly evoked in this literature, the effect of labels is unlikely to work through changing beliefs about CO_2 emissions, but may, instead, work through changing the salience or perceived social norms of climate-friendly behavior.

The literature on measuring misperceptions features a number of papers that elicit broad knowledge of the climate change phenomenon and link it to measures of concern and policy support (Tobler, Visschers and Siegrist, 2012; Shi et al., 2016; Klenert et al., 2018; Dechezleprêtre et al., 2022; Fairbrother, 2022). Attari et al. (2010) find that people underestimate the energy use associated with different activities. Closest to our paper, Camilleri et al. (2019) elicit perceptions of greenhouse gas emissions associated with the production and transportation of food and the use of several electric appliances. Participants underestimate emissions for all products and activities, but especially those in the food domain.

We go beyond eliciting unincentivized point estimates of carbon impact, by administering incentivized elicitations of belief distributions and combining them with revealed preferences over mitigation[1] in a structural model. Our approach has the potential to overturn conclusions

[1]There is a large literature on willingness to pay to reduce climate impact, often using unincentivized surveys and contingent valuation methods (see Nemet and Johnson (2010) for a review) and the literature on willingness to pay to reduce emissions from specific sources like car transport (Hulshof and Mulder, 2020) or flights (Bernard, Tzamourani and Weber, 2022). Two recent studies use incentivized revealed preference techniques to elicit WTP for a single emissions amount. Löschel, Sturm and Vogt (2013) find an average WTP to buy emissions offsets for one ton of CO_2 of 12€, whereas Diederich and Goeschl (2014) find a mean of 6.30€. Andre et al. (2021) show that the willingness to donate to a charity to fight climate change is affected by perceived social norms.

about the optimal targeting of information that are derived solely on the basis of point estimates. Yet, we find that the predictions of our structural model are broadly in line with results in Camilleri et al. (2019) and with the results we would have obtained looking only at beliefs. Therefore, our representative survey and structural model lend robustness to established results. However, our experiment then demonstrates that the presence of misperceptions does not imply that correcting them yields behavioral change. More generally, our results reveal limits of survey evidence in guiding policies related to voluntary climate change mitigation. We find that economic primitives such as the valuation of carbon emissions and beliefs about their size, measured with state-of-the-art elicitation techniques, have little predictive power over consumer decisions in our experiment.

The literature on labeling studies the effect of climate labels that code high and low-impact consumption in easily digestible ways (see Taufique et al. (2022) for a summary). While most studies in this literature focus on hypothetical choices, several papers have looked at real consumption choices in the context of restaurants or university canteens, sometimes studying labels in combination with another information intervention, like posters (e.g., Spaargaren et al., 2013; Visschers and Siegrist, 2015; Brunner et al., 2018; Soregaroli et al., 2021; Lohmann et al., 2022). Other studies have provided shoppers in (online) supermarkets with informative labels about specific products or shopping baskets (Vlaeminck, Jiang and Vranken, 2014; Elofsson et al., 2016; Perino, Panzone and Swanson, 2014; Kanay et al., 2021; Bilén, 2022), or informed them via a cell phone app (Fosgaard, Pizzo and Sadoff, 2021). Most of these papers find a small and short-lived effect of labels on behavior and ultimate emissions.[2] However, null results have been reported for specific products like detergents (Kortelainen, Raychaudhuri and Roussillon, 2016). Some studies look at the effect of labels on meat consumption specifically. Camilleri et al. (2019) conduct an experiment where participants were asked to purchase a can of soup. Participants were less likely to buy high-carbon-impact beef soup when a GHG impact label was present. Bilén (2022) finds suggestive evidence that when carbon labels are introduced in a supermarket, customers reduce their purchases of beef.[3]

The fact that most studies on carbon labels find a small positive effect on green consumption, while our information treatment yields a well-powered null effect is instructive. Our experiment is designed to study the effect of changing beliefs about CO_2 emissions on climate-friendly consumption while keeping the salience of climate change constant. Instead, the introduction of climate labels in the above studies may have yielded behavior change by increasing the salience of climate change or by changing the perceived social norms, channels we rule out.

Understanding the channel through which information can change behavior is more than a

[2]Labeling has also been shown to affect energy saving (Allcott and Taubinsky, 2015). However, it is unclear whether these results extend to CO_2 emissions, since energy costs are paid by the consumer, but emissions are not. Indeed, in an experiment concerning a hypothetical choice for a water heater, Newell and Siikamäki (2014) show that CO_2 emission information is less effective in inducing sustainable choices than informing people about energy costs.

[3]Moreover, a review by Bianchi et al. (2018) finds that information can affect intentions to buy meat. Carlsson, Kataria and Lampi (2022) finds substantial resistance to switching away from meat among Swedish consumers. Jalil, Tasoff and Bustamante (2020) show that a 50-minute lecture on meat consumption reduces purchases of the meat-based meal at the university canteen.

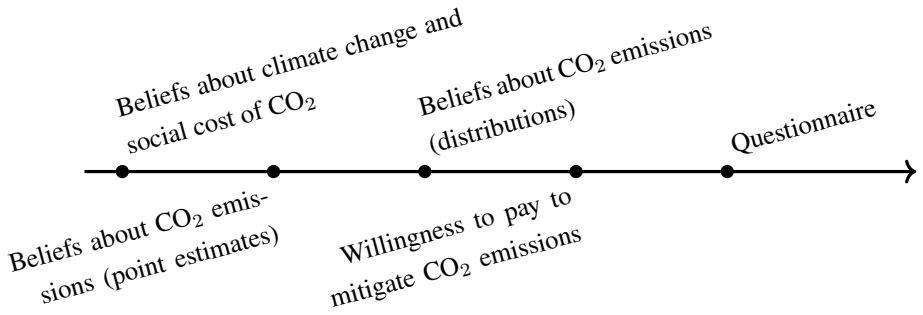


Figure 4.1: Timeline of the climate survey.

theoretical curiosity: it matters for policy. If beliefs are key, then education can play an important role. If the effect of information is driven by salience or perceived social norms, then policy makers will have to design interventions that effectively change the attentional, emotional, or social context at the point of purchase. We provide clean evidence that beliefs about CO_2 emissions are of second-order importance in driving green consumption. An immediate corollary of this result is that much remains to be understood about why and when information and labels actually affect consumption behavior.

4.2 Climate Survey

Our initial survey measures consumers' existing beliefs about CO_2 emissions generated in the production of common consumer goods, as well as their willingness to pay (WTP) to avoid CO_2 emissions. These quantities subsequently serve as inputs for a structural model that allows us to make predictions about the provision of information, as we explain in Section 4.3. Figure 4.1 shows the four tasks that constitute the survey. The first task asked general questions about climate change facts and the social cost of carbon. The next two tasks focused on eliciting beliefs, where we collect both point beliefs and belief distributions of CO_2 emissions from several common consumer products and activities. The last task elicited willingness to pay for mitigating CO_2 emissions.[4] After participants completed all four tasks, we asked them about their demographics and revisited the products and activities from tasks two and three to ask them about their consumption frequency in these categories.

Our elicitation methods used incentive-compatible payment schemes developed in the experimental economics literature, while keeping the instructions and the interface as simple and participant-friendly as possible to allow for a representative sample to take part. Below we elaborate on each of the elicitation procedures in more detail. Online Appendix D.1.1 contains additional information about the steps we took to maximize the data quality.

[4]The survey had one additional part that we analyze in a separate paper. At the end of the survey, we provided subjects with information about the actual impact of a subset of the product list (three or six randomly selected products). We then re-invited the subjects two weeks later to test their recollection of this information.

Table 4.1: List of consumer products and actions.

		Emission size		
	Quantity	Estimate	Unit	Source
Beer	12 fl oz	1.46	mile	Poore and Nemecek (2018)
Phone call	1 hour	1.55	mile	Smith et al. (2013)
Microwave	1000W, 2 hour	1.76	mile	UK BEIS (2020)
Milk	1 cup	2.60	mile	Poore and Nemecek (2018)
Egg	6 eggs	4.81	mile	Poore and Nemecek (2018)
Poultry meat	7 oz	6.78	mile	Poore and Nemecek (2018)
Shower	Average usage	3.90	mile	Hackett and Gray (2009)
Dark chocolate	100g	16.03	mile	Poore and Nemecek (2018)
Coffee beans	1 lb	44.41	mile	Poore and Nemecek (2018)
Beef	7 oz	68.39	mile	Poore and Nemecek (2018)
Flight	SFO to LAX	304.60	mile	UK BEIS (2020)
Gas heating	One month	606.68	mile	Padgett et al. (2008)
Car	drive 1 mile	291.00	gram	UK BEIS (2020)

Belief elicitation

At the start of the survey, we elicited participants' beliefs about the CO_2 emissions generated by driving one mile by car. We then elicited beliefs about 12 common consumer products and activities listed in Table 4.1. We included food items, the use of household appliances, and transportation. We provided participants with information about the product specification and the type of emissions we considered. Table 4.1 presents the scientific estimates we used to incentivize the guesses together with their source.[5] We took these estimates from top-tier academic journals or from the estimates the UK government uses for its environmental regulations. We disclose these scientific sources only at the end of the experiment.

To make the answers more meaningful to subjects, we did not elicit emissions in grams, but asked about the number of miles by car one needs to drive to emit as much CO_2 as the product in question, an approach in line with previous studies (Camilleri et al., 2019). Since we also elicited the conversion from a mile driven by a car to grams of CO_2, we can convert all measures to the perceived grams equivalent (see Table D.1.3 and Figure D.1.6 in the Online Appendix). Moreover, the model we describe in Section 4.3 further mitigates any concern that systematic misperceptions about the CO_2 emissions associated with driving bias our predictions, because these predictions will be independent of the denomination of CO_2 emissions.

We divided the belief elicitation into two parts. We first elicited a point estimate for the modal value of the emissions. Participants indicated how much CO_2 each of the 12 products in Table 4.1 emitted relative to driving one mile by car. Participants answered all 12 questions on one page, and the order of the products was randomized across participants (Figure 4.2A). In the rest of the experiment, the same order was used every time participants answered additional

[5]Participants could learn the detail of what the scientific source took into account in calculating the size of CO_2 emissions. See Table D.1.1 in the Online Appendix.

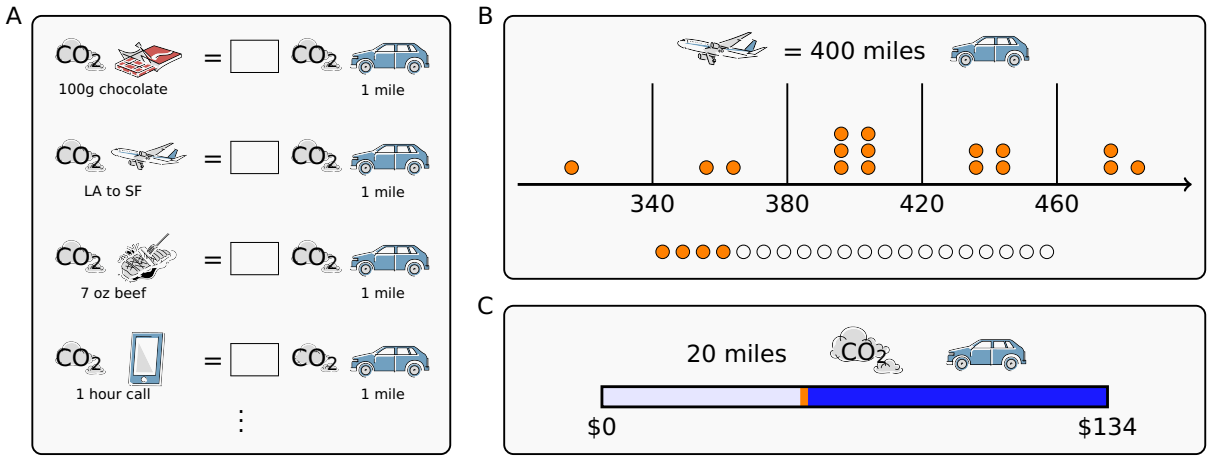


Figure 4.2: Illustration of the belief and WTM elicitation interface. (A) Point-belief elicitation task. (B) Bins-and-balls belief elicitation task. (C) WTM elicitation task. *Notes*: Panel B shows an example in which a participant stated 400 in the previous point belief elicitation task and is now asked to allocate 20 balls into five bins, centered around this number. See Online Appendix D.1.1 for screenshots of the interface.

questions about these products. To help participants keep track of their guesses and the rankings of the products, we presented an interactive box summarizing their (current) answers at the bottom of the page, including the ranking of the products by estimated impact. We incentivized a correct point estimate with a $5.36 (£4) bonus. We considered an estimate correct if it was within a 5% interval from the scientific estimate. This incentive scheme truthfully elicits the mode of the subjective probability distribution about the scientific estimate (Schlag, Tremewan and van der Weele, 2015).[6]

In order to understand the participants' confidence in their answers, we then elicited the subjective probability distribution about the size of CO_2 emissions. For each product, we presented five "bins" around the point estimate participant reported in the first part and asked the participant to allocate 20 balls into these five bins. We told participants that each bin represents an interval that might contain the scientific estimate and that they should allocate the balls to represent their level of confidence that the estimate is in fact in that bin. Figure 4.2B provides an illustration. We incentivized the elicitation by randomly selecting one of the bins, and scoring the answer according to a randomized quadratic scoring rule. This mechanism encourages participants to truthfully reveal their belief that the scientific estimate falls in a particular bin (Schlag and van der Weele, 2013). To keep things simple and avoid information overload, we did not provide participants with the exact details of the scoring rule, which were available with a mouse click, but told them that they would maximize their expected earnings by answering truthfully, an approach suggested by Danz, Vesterlund and Wilson (2022).

[6]We did not incentivize the questions about the CO_2 emissions and the social cost of driving one mile by a car as we realized that answers to these questions can be straightforwardly obtained on Google.

Willingness to mitigate

After the belief tasks, we elicited the participants' willingness to pay for mitigating CO_2 emissions of different sizes. We call this measure *willingness to mitigate (WTM)*. To introduce real consequences in the elicitation task, we offered participants trade-offs between monetary payments and carbon offset certificates. More precisely, we used donations to Carbonfund.org (`https://carbonfund.org/`), a charity that finances various projects to offset CO_2 emissions and offsets one ton of CO_2 for every $10 donated.

To cover the amounts of the emissions generated by all the consumer products we asked in the survey, we elicited the WTM for eight levels of CO_2 emissions, corresponding to emissions generated by driving 1, 5, 20, 50, 100, 200, 450, and 700 miles by car. Participants expressed their WTM to offset these amounts of CO_2 using a slider between $0 and $134 (£100), see Figure 4.2C.[7] The interface was designed to help participants make consistent choices and avoid anchoring. To this end, the sliders for each emission quantity were all displayed on the same screen, and the bottom of the page featured a graphical summary of reported WTMs by emission quantity (see Online Appendix D.1.1).

We incentivized the WTM with a Becker–DeGroot–Marschak (BDM) mechanism, which means that reporting the true WTM is in the best interest of the participant.[8] To make sure our donations were credible to participants, we emphasized that our ethics committee does not allow misleading instructions, and promised to send them the carbon offset receipts from the experiment. The method above provides data that are censored at $134. To mitigate this problem, we added a second, unincentivized set of questions. For every emission level for which a participant reported a WTM of $134, we asked the participant to indicate for which amount of money he or she would have agreed to allow the emissions. The participant could either type in a number or check a box to signal that no monetary compensation would have been enough.

At the end of the session, we asked a series of questions about demographic background, consumption habits (about the 12 products), and attitudes toward climate change. See Online Appendix D.1.1 for the complete list of questions.

Implementation

We recruited 1,430 participants on Prolific (`https://www.prolific.co/`) between the 3rd and 6th December 2020, and 1,022 of those completed the whole survey.[9] We restricted participation to US residents and we aimed to collect a sample representative for age, gender,

[7] Participants could also express their WTMs either in GBP (between £0 and £100), the official currency of Prolific, or in USD (between $0 and $134).

[8] We randomly selected one number from a discrete set of values between 0 and 100. If the number was bigger than what the participant reported, we paid the participant a bonus equal to the randomly selected number. If, instead, the number was smaller than the participant's report, we donated to Carbonfund.org as much money as needed to compensate for the CO_2 emissions stated in the question.

[9] We ran extra sessions on 21st and 22nd December 2020 to recover some participants' demographic data. These data were not originally saved due to a failure in the survey code. We managed to retrieve the data of 67 of the 69 participants for which the failure was verified. Only the demographic questions were asked in these extra sessions.

and ethnicity.[10] Our sample is on average 42.7 years old (SD $= 15.4$) and 48.3% of the participants identified themselves as male. Table D.1.2 in the Online Appendix shows the demographic characteristics of the sample.

To make the instructions as accessible as possible, we used slides that displayed the instructions step by step with explanatory images complementing the written text. Besides, we divided the instructions into 5 blocks. After each block, we asked participants to answer several comprehension questions. We did not allow subjects to continue with the experiment until they answered all the questions of each block correctly. In total, participants had to answer 21 comprehension questions.

At the end of the experiment and for every participant, we randomly selected one question from the entire study. Depending on the participant's answer to that question and luck, we paid them a bonus. This incentive mechanism elicits truthful answers in experiments with multiple tasks (Azrieli, Chambers and Healy, 2018).[11] Participants received \$10.05 for completing the study plus a variable bonus depending on their answers (mean $= \$2.67$, SD $= 4.31$).[12] The median survey completion time was 55 minutes.

4.2.1 Results

Beliefs. Participants estimated CO_2 emissions from 12 common consumer products and activities in terms of miles of driving by car. Table 4.2 shows summary statistics of reported (point) beliefs and Figure 4.3A plots them against scientific estimates of CO_2 emissions.[13,14] Median beliefs lie below the identity line for all but one (microwave) products, indicating that participants underestimated the size of CO_2 emissions. This is in line with findings in Camilleri et al. (2019), despite differences in the sets of products, elicitation methods, and the reference items (lightbulb vs. car).

The fraction of participants who underestimated the size of emissions varies from 41% (microwave) to 92% (gas heating), with this fraction increasing in the true size of the emissions. Flying is a notable exception to this trend: it is a highly polluting activity but its emissions are underestimated only by 59% of participants. This could be due to the ample coverage of emissions from flying from media outlets, or because subjects simply took as an estimate the

[10] We noticed that participants in the oldest age bracket (above 58 years old) particularly struggled with the comprehension questions about the WTM resulting in many dropouts on the page where those questions were asked. As subjects in this demographic category were hard to recruit, we opted to give them a second chance to complete the experiment. On 7th December 2020, we invited them to re-start the experiment from the WTM instructions and we gave them the solutions to two of the 7 related comprehension questions. Of the 41 subjects that were allowed to restart the experiment, 22 completed it.

[11] The probability with which a question was selected for payment was not uniform but depended on the part of the experiment that the question came from. In the instructions, we informed participants of the probability that the question was drawn from each of the different tasks of the experiment.

[12] Participants received the completion reward and the bonus only if they completed the second part of the experiment. This second part of the experiment took place two weeks after the first. Participants that completed both parts of the experiment received a total completion reward of £10 and an average bonus of £2.20. Following the participants' decisions in the experiment, we donated \$88 to Carbonfund.org, offsetting 8.8 tons of CO_2 emissions.

[13] We focus on median beliefs since there are several extreme outliers.

[14] Figure D.1.5 in the Online Appendix shows empirical CDFs of reported CO_2 emission sizes for each product.

Table 4.2: Summary statistics of elicited (point) beliefs about CO_2 emissions from 12 consumer products and activities.

			Belief			
Product	Emissions	Unit	Q1	Median	Q3	Under-est.
Beer	1.46	miles	0.50	1.20	6.00	0.516
Phone call	1.55	miles	0.40	1.00	5.00	0.549
Microwave	1.76	miles	0.80	2.15	10.00	0.406
Milk	2.60	miles	0.50	2.00	8.00	0.570
Shower	3.90	miles	0.50	1.50	5.00	0.689
Egg	4.81	miles	0.50	1.50	6.00	0.697
Poultry	6.78	miles	0.60	2.50	10.00	0.676
Chocolate	16.03	miles	0.40	1.20	8.00	0.831
Coffee	44.41	miles	0.50	2.00	10.00	0.885
Beef	68.39	miles	1.00	5.00	20.00	0.858
Flight	304.60	miles	10.00	150.00	600.00	0.586
Gas heating	606.68	miles	3.00	20.00	100.00	0.919
Car (drive 1 mile)	291.00	grams	5.03	85.00	403.00	0.677

Notes: The last column "Under-est." shows the fraction of participants who underestimated the size of emissions.

driving distance between San Francisco and Los Angeles (≈ 350 miles), which is close to the right answer.

Even though participants misperceived the size of CO_2 emissions from each product, they had a good understanding of which products emit more CO_2. As Figure 4.3B shows, the "true" ranking of emission sizes based on scientific estimates and the ranking "revealed" by each participant's estimate are positively correlated.[15]

All the qualitative results of this section replicate if we express participants' beliefs in terms of grams of CO_2 using their beliefs about the CO_2 emissions linked with driving one mile by car. Figure D.1.6 in the Online Appendix shows that, since participants underestimate the grams of CO_2 emitted when driving, the underestimation is more severe if we express the beliefs in grams.

Taken together, the belief elicitation tasks in the climate survey suggest that consumers significantly underestimate the size of CO_2 emissions associated with common consumer products and activities, but they have more accurate perceptions about the ordinal ranking of CO_2 emissions.

Willingness to mitigate. We now turn to participants' willingness to mitigate CO_2 emissions. Note that we elicited WTM for eight levels of CO_2 emissions, that correspond to emissions generated by driving 1, 5, 20, 50, 100, 200, 450, and 700 miles by car. On average, participants

[15]We also calculated Spearman's rank-order correlation between the actual ranking of CO_2 emissions and "revealed" ranking of emissions for each participant. About 95% of the participants exhibited a positive correlation, and 45.6% of the participants exhibited a statistically significant positive correlation (two-sided, $p < 0.05$). The average correlation coefficient is $\rho = 0.559$.

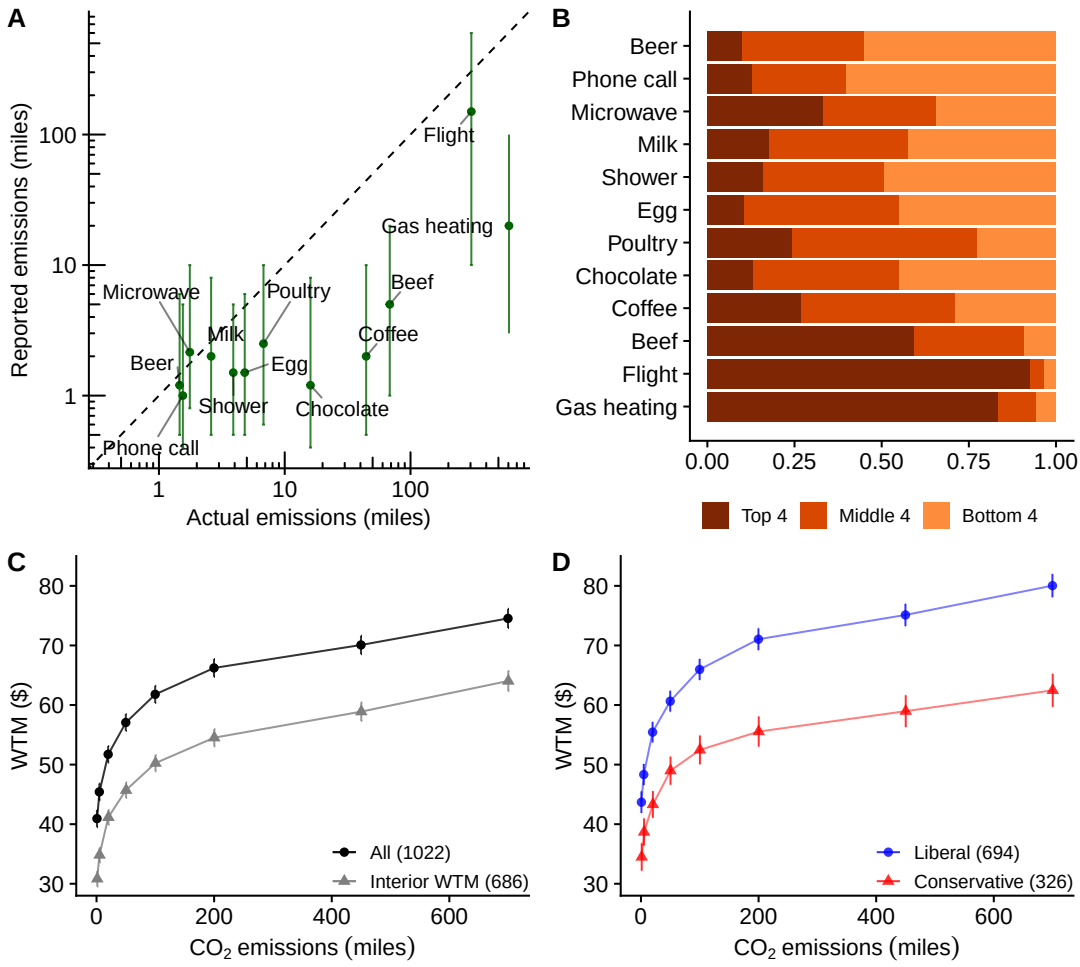


Figure 4.3: Beliefs and willingness to mitigate. (A) Summary statistics of reported CO_2 emissions (median and IQR). Axes are on a logarithmic scale. (B) Ranking of reported emission sizes. Products are sorted by the true emission size from low to high. (C) Concave WTM (mean and SEM). (D) WTM and political view (mean and SEM). *Notes*: In panels C and D, numbers in parentheses indicate the number of observations. In panel D, "somewhat liberal" and "somewhat conservative" are grouped into liberal and conservative, respectively.

have positive and sizable WTM for all levels of CO_2 emissions, and they exhibit a concave pattern (Figure 4.3C, dark line). Moving from emissions equivalent to driving 5 miles to 20 miles, a four-fold growth, increases the WTM by \$6.3 on average, while moving from 5 to 200 miles, a jump 10 times as large as the previous one, pushes the average WTM by only \$20.8. The marginal willingness to pay for mitigation decreases as the emission size increases, confirming findings in Pace and van der Weele (2020). This pattern is not due to top-censoring at \$134 — the concave pattern is preserved even when we focus on 686 participants whose WTMs are all strictly between \$0 and \$134 (Figure 4.3C, light gray line). See Tables D.1.4 and D.1.5 in the Online Appendix for summary statistics of WTMs and the number of "corner" observations for each level of emissions.

As in elicited beliefs, we observe strong correlations between WTM and some of the demographic characteristics. Participants who identified themselves as liberal on the political spectrum have uniformly higher WTM than conservative participants (Figure 4.3D). Female

participants have higher WTM than male participants, and participants in the age ranges of 18-37 and 58 and older have higher WTM than those between 38 and 57 years of age (Figures D.1.7 and D.1.8 in the Online Appendix).

Figure 4.3C shows a smooth and concave WTM curve at the aggregate level, but it masks substantial heterogeneity across participants. There are 52 participants who "do not care" about CO_2 emissions and request \$0 for all eight levels of emissions, and there are 77 participants who are "deontological" and request \$134 all the time. We can classify the shape of the WTM curve. We observe that 31% of individual-level WTM curves are concave, and 28% of WTM curves are non-monotonic. Less than 10 are convex. There are only 44 cases of decreasing WTM curve, an irrational pattern of WTM that is not captured by small mistakes. See Online Appendix D.1.2 for details.

In the next section, we describe how to combine these measures for the prediction of information provision.

4.3 Modelling the impact of information

In this section, we outline a simple formal framework to combine beliefs about the impact and willingness to mitigate and produce a prediction about the resulting consumer decision. The key assumption is that consumers suffer a cost from the expected emissions produced by their actions and that they make utility-maximizing decisions about the quantities of emissions. Our approach is inspired by findings that subjects make rationalizable trade-offs about payoffs for themselves and others that allow for the construction of a utility function (Andreoni and Miller, 2002; Fisman, Kariv and Markovits, 2007).

Consider a consumer who gets material utility v from purchasing a good or activity. We assume that the good (or activity) is sold at a market price of p and is associated with a quantity of CO_2 emissions $c \geq 0$. The consumer's utility from consuming the product is:

$$U = v - p - w(c),$$

where $w : \mathbb{R}_+ \to \mathbb{R}_+$ captures the psychological cost from CO_2 emissions. We assume w is strictly increasing and $w(0) = 0$.

In writing the preferences in this way, we are making two assumptions. First, for simplicity, we assume that the consumer's overall utility is additively separable in v and in the psychological cost of emitting CO_2. Second, we assume that the psychological cost only depends on the emissions associated with the current purchase and not on the emissions linked to previous consumption of the same or other products. This last assumption finds support in our willingness to mitigate data. For us to observe the concavity of the function w, it must be the case that the consumers consider the emissions they can offset in the experiment separately from all the emissions they have generated so far. Without this "narrow bracketing" of emissions, participants with a concave WTM would report a flat WTM curve in the survey.[16]

[16]Narrow bracketing has also been documented in choices over monetary outcomes (Rabin and Weizsäcker,

We assume that the consumer may not have precise knowledge about emission sizes c, but has some beliefs about them. Let F denote her belief about c. With this subjective belief and following standard expected utility, the consumer's preferences can be expressed as

$$U = v - p - E_F[w(c)].$$

Two key ingredients in this framework are the function w capturing psychological cost and the subjective belief about CO_2 emissions F. The climate survey we discussed above is designed to measure these two quantities as precisely as possible. Remember that we used "miles driving a car" as the common unit of emission size in the belief and WTM elicitation tasks in the survey.

The WTMs stated by each participant provide information about w. Requesting a bonus of y_m to allow emitting CO_2 corresponding to emissions generated by driving m miles by a car, c_m, reveals

$$y_m = w(c_m),$$

assuming a linear utility for money. Using eight pairs of observed (c_m, y_m) and extrapolating (see Online Appendix D.1.3), we can recover w for each participant. Hereafter we will refer to w as the WTM function.

Similarly, we use the second part of the belief elicitation, the bins-and-balls task, to recover subjective belief *distribution* F_k for each product k. See Online Appendix D.1.3 for details.

Quantifying the effect of information

Given a WTM function w and a subjective belief distribution F about CO_2 emissions associated with a good or activity, we can calculate the *expected WTM*,

$$\overline{W}(w, F) = E_F[w(c)] = \int w(c)\mathrm{d}F(c).$$

This quantity captures the extra amount of money a consumer is willing to pay in order to consume an imaginary, "carbon-neutral," version of the good or activity, taking into account the lack of knowledge about the actual size of CO_2 emissions.

We model an *information policy* as a device that shifts consumer i's belief about CO_2 emissions associated with good k from F_{ik} to F_k^*, a degenerate distribution at the "true" size of CO_2 emissions.[17] The difference in expected WTM before and after information for each consumer i and product k is given by

$$\Delta_{ik} = \overline{W}(w_i, F_k^*) - \overline{W}(w_i, F_{ik}).$$

2009; Ellis and Freeman, 2020) and in work choices (Fallucchi and Kaufmann, 2021). The concavity of the WTM function also implies that narrow bracketing is essential for an information campaign to have any effect on behavior. Given the beliefs and consumption levels of the average US consumer, broad bracketing implies that they will be on a flat part of w.

[17] Note that we impose an assumption that the consumer trusts the information and fully updates her belief, but the framework can easily accommodate the possibility that the updated belief is not exactly F_k^*, reflecting the idea that the consumer has some doubt in the information or has difficulty in giving up her original belief.

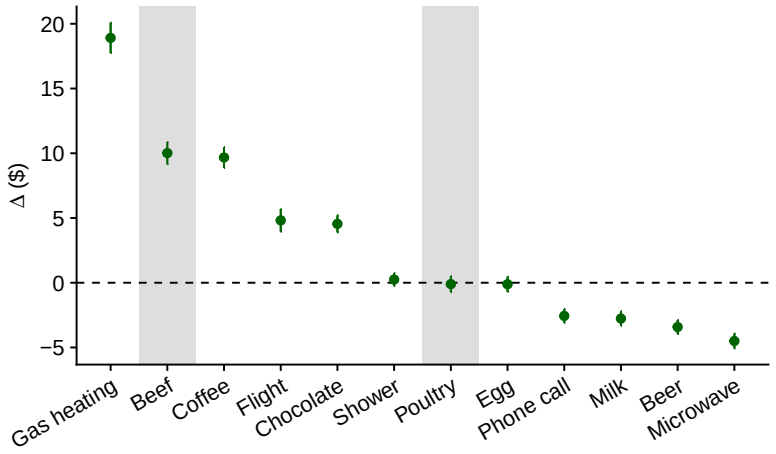


Figure 4.4: Predicted effect of information provision Δ_k for each product. *Notes*: The reference group G is the entire sample of survey participants ($N = 1,022$). Bars indicate SEM.

If $\Delta_{ik} > 0$, information raises the psychological cost from consuming a unit of good k for consumer i through a change in her beliefs. If this increase is large enough, information may result in a change in consumer i's buying behavior.

Finally, we define the effect of information provision on the consumption of good k, Δ_k, as the sample average of Δ_{ik} with respect to a reference group of agents G:

$$\Delta_k = \frac{1}{|G|} \sum_{i \in G} \left(\overline{W}(w_i, F_k^*) - \overline{W}(w_i, F_{ik}) \right).$$

Again, if $\Delta_k > 0$ and demand is downward sloping, then information is predicted to result in a decrease in buying behavior in target group G.

Several features of our structural model bear mentioning. First, the effect of an information campaign Δ_k has a simple interpretation: providing accurate information on the CO_2 emissions of product k increases the average subjective cost of consuming product k by Δ_k dollars. Therefore, Δ_k can be thought of as the equivalent of a price increase. As with a price increase, the ultimate effect of information on consumption choices will be mediated by a product's elasticity of demand, something we will address in the next section.

Second, because our model combines beliefs and willingness to mitigate CO_2 emissions that were both expressed as miles-driven-in-a-car equivalents, the unit of denomination of CO_2 emissions drops out of our prediction. This allows us to use an intuitive and common way of denominating CO_2 emissions while assuring that any systematic misperceptions about the climate impact of driving do not affect our predictions.

Prediction

We now calculate our measure of the effect of information provision using the data from the survey. Taking the entire sample of 1,022 participants as the reference group G, we obtain Δ_k for each product k as shown in Figure 4.4.

We observe a substantial variation in the effect of information provision. We expect a positive effect for five products (gas heating, beef, coffee, flight, chocolate), no effect for three products (shower, poultry, egg), and a negative effect for four products (phone call, milk, beer, microwave). Note that we expect a larger effect of information for products with larger CO_2 emissions: the ordering in Figure 4.4 is almost the mirror image of the ordering in Table 4.1. This is because the fraction of participants who underestimates the size of emissions is larger for these products, and our measure favors these participants as long as their WTM function responds to the size of emission (i.e., w is not constant on the relevant range). These predictions have received some support in the empirical literature. For instance, the negative effect for electrical appliances has been documented in several empirical papers (Rodemeier and Löschel, 2020; d'Adda, Gao and Tavoni, 2022). In a labeling intervention in an online Swedish supermarket, Bilén (2022) observes an effect for beef, but not poultry.

Taking different subgroups of participants as the reference group G, we can also quantify Δ_k depending on the target population. Figure 4.5 conducts such an exercise, focusing on two meat products, beef and poultry, that will be the subject of the experiment in the next section. While panel A shows the aggregate effect, panels B-G disaggregate the predictions across several subgroups. These panels illustrate the advantages of integrating preferences and beliefs over simpler approaches, like simply targeting populations with a high willingness to pay. For instance, the model predicts a larger effect for males than females (panel C), and for participants who have conservative political views than those with liberal views (panel D), despite the fact that in both cases, the former group has a lower WTM (see Figure D.1.7 in the Online Appendix). The reason is that these groups also have larger underestimations of climate impact, which more than offsets their lower WTM, resulting in a higher predicted impact of information.

Moreover, we can assess the robustness of our model's prediction for beef consumption. The predicted effect of an information campaign may be interpreted as a "subjective price increase" of the product under investigation. Just like with a conventional change in prices, a price increase will have little effect on demand if it is primarily experienced by individuals whose demand is inelastic or by individuals who do not consume the product, to begin with. Thus, one might ask whether the effect differs between groups that might have different elasticities of demand, based on self-reported consumption patterns in the survey.

Such an exercise is shown in panels E-G of Figure Figure 4.5. The predicted effect of an information campaign is higher for those who are more prepared to reduce future meat consumption in light of its CO_2 emissions (panel E), those who find it "not difficult" to reduce beef consumption and hence should have more elastic demand (panel F), and those who consume beef below the median frequency (panel G). However, in each case, the effects of these splits are relatively small, illustrating that our predictions about interventions for beef are robust to prevailing demand levels and elasticities.[18]

[18]The prediction is based on the consumption of 7 oz of beef and poultry, the size of meat products participants reported their beliefs about CO_2 emissions. Figure D.1.10 in the Online Appendix shows the prediction about 5 lb (80 oz) of beef and poultry, the size of meat products offered to participants in the Meat Experiment, by "scaling up" their belief distributions by the factor of $80/7$, which shifts Δ_k upward for both products. The overall prediction is different in absolute terms (e.g., the bottom panel of Figure D.1.10 shows a positive overall effect of information

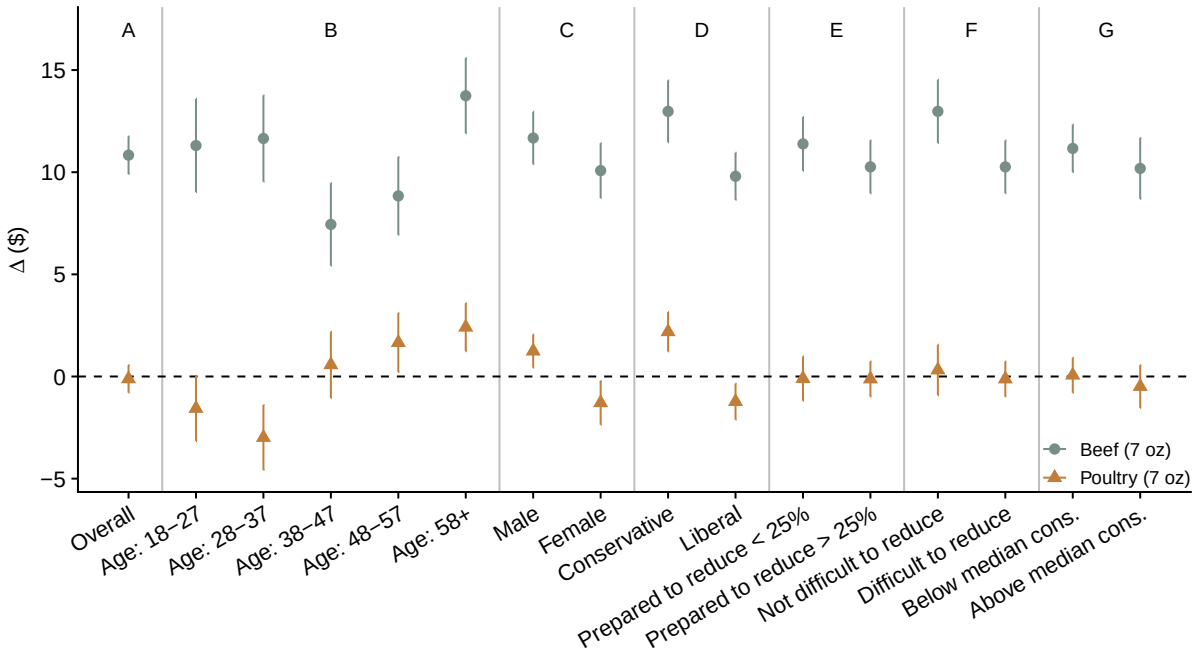


Figure 4.5: Predicted effect of information provision for each demographic group. *Notes*: (D) "Somewhat liberal" and "somewhat conservative" are grouped into liberal and conservative, respectively. (E) "Are you prepared to reduce your future consumption of beef/poultry in light of its CO_2 emission footprint?" (F) "How difficult would it be to reduce your current consumption of beef/poultry by half?" (G) "How many times do you eat beef/poultry per week?" Bars indicate SEM.

4.4 Meat Experiment

We now turn to test the predictions we derive from our calibrated structural model in Section 4.3. To this end, we compare the effect of information between beef and poultry meat. There are three main reasons for choosing these two products. First, meat products are an important application, as meat (and especially beef) consumption makes a meaningful contribution to climate change and is one of the main sources of emissions that are under the direct control of consumers.[19] Second, these two products are comparable in many respects as they fall into the same food category and may be considered substitutes for certain purposes. Third, despite their similarity, these two products have very different predicted effects of information provision, as we show in Figure 4.4. While the predicted effect of information on beef consumption is among the very highest on our product list, it is approximately zero for poultry. This is mainly because beef production is about 10 times more carbon-intensive than poultry production, an effect that is not incorporated into the expectations of consumers, and hence subject to correction through information provision.[20]

even for poultry), but qualitatively the results do not change: the model still predicts larger effects of information for beef.

[19] Alexandre Koberle, Grantham Institute for Climate Change, Imperial College London, writes that "Next to flying less, it is probably right to say that, as individuals, reducing beef consumption is the most significant contribution directly under our control" (Vetter, 2020).

[20] This difference results mainly because beef involves the release of large amounts of methane, a greenhouse gas with about 30 times the warming equivalent of CO_2, and because beef requires large amounts of feed which spurs deforestation (Poore and Nemecek, 2018).

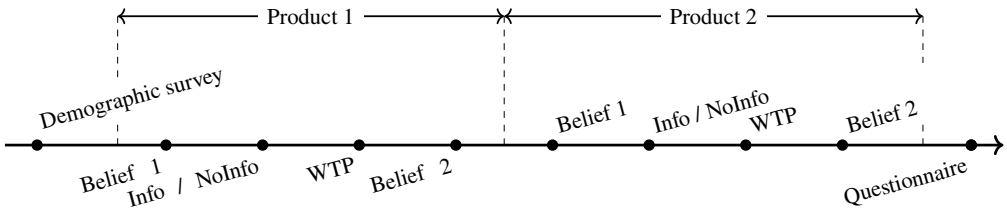


Figure 4.6: Timeline of the meat experiment.

Thus, the main hypothesis that we test in our experiment is that information provision about carbon impact will have a bigger impact on consumer valuations of beef products than on valuations of chicken products.

Design

In this experiment, we offered participants an opportunity to purchase a bundle of high-quality meat products, either 10 beef sirloin steaks or 10 skinless chicken breasts. We kept the features of bundles as close as possible: they were sold on a premium online butcher Porter Road (`https://porterroad.com/`); they weighed about 5 lb ($\approx$ 2.3 kg); they cost \$100 (at the time of designing the experiment in 2021); they were pasture-raised in the US without hormones and antibiotics. We provided these descriptions in the relevant part of the instructions.

Across treatments, we varied between subjects whether the participants received information on the CO_2 emissions associated with beef and poultry meat (Info treatment) or not (NoInfo treatment). In keeping with our climate survey, we provided the information in terms of the number of miles by car one needs to drive to emit as much as 1 lb of the meat. We pinned down participants' beliefs about the car CO_2 emissions by including a scientific estimate of these emissions (in ounces) in the instructions. In this way, we made sure that our information treatment could only impact the beliefs about the meat. The information about car emissions was available in all treatments.

As an additional manipulation, we varied whether the participants were first offered the beef bundle (BeefFirst treatment) or the poultry bundle (PoultryFirst treatment). For these two products, subjects remained in the same information treatment. We test our main hypothesis about the differential impact of information for the two products using the first product offered in the experiment. The second part allows us to evaluate spill-over effects, whereby information about beef affects beliefs and WTP for poultry or vice versa.

The timeline of the experiment is illustrated in Figure 4.6. The experiment has two parts, one per each of the products we offer. The two parts followed the same structure. Each part of the experiment started with a description of the bundle the participants could purchase as well as its retail value (\$100). We then asked the participants to guess the average CO_2 emissions associated with the production and distribution of 1 lb of the type of meat that they were offered. As in the climate survey, participants expressed their guesses in terms of CO_2 emitted by driving

one mile by car.[21]

To help participants to get a sense of the magnitudes of emissions, just before they could express their guesses, we informed them of how the CO_2 emissions from driving one mile by car compared with the emissions generated by the production and distribution of 12 fl oz of beer and by taking a plane from Los Angeles to San Francisco. We provided this baseline information to all participants to keep the salience of emissions and possible norms around low-carbon consumption constant across treatments. To incentivize belief elicitation, we used the same sources of scientific estimates as in the climate survey and we rewarded accurate guesses (those within $\pm 5\%$ of the scientific estimate) with a \$0.5 bonus.

Next, we had our treatment manipulation. The participants in the Info treatments were informed about the average emissions associated with the meat product they could purchase. To make sure that the participants paid attention to the information, we asked them to identify the true size of the emissions among three possible options. The participants in the NoInfo treatments, instead, saw three random numbers and answered a similar question.[22]

We then elicited participants' WTP using a two-stage multiple price list (MPL) with forced single switching.[23] On the first list, participants saw 11 choices between two options: the left option is the meat bundle and the right option is the monetary bonus ranging from \$0 to \$100 in \$10 increment. In the remainder, we refer to this bonus as the "price", although it was not framed as such in the experiment. The second list "zoomed in" around the switching point and asked another nine questions. With this procedure, we measured WTP in the precision of \$1.[24] The instructions encouraged the participants to think about their own valuation of the meat bundle and to use this valuation to make the decisions.

After completing the MPL task, we asked participants to guess one more time the size of the emissions associated with the meat product they had the opportunity to purchase. This second guess was not incentivized.

The second part of the experiment followed the same structure as the first one, but it asked participants about their beliefs and WTP for the other meat bundle—the poultry bundle if the first part was about beef, and the beef bundle if the first part was about poultry. Thus, in the Info treatments, participants saw the information about the CO_2 emissions associated with the new meat bundle together with all the information previously provided. In the NoInfo treatment, instead, participants saw four randomly generated numbers.

[21] We did not elicit belief distributions to fit the survey in the time constraint of 15-20 minutes.

[22] In both treatments, participants were allowed to proceed regardless of their answers. However, participants who answered incorrectly received an alert warning them of the mistake and repeating the correct answer.

[23] We used an MPL instead of the slider interface from the climate survey since we elicited only two valuations in this experiment while we elicited eight in the survey. The small number of valuations makes an elicitation strategy that requires simpler instructions (MPL) preferable to a strategy that requires more complicated instructions but allows the participants to input their decisions more quickly.

[24] We used a BDM procedure to make this two-stage MPL incentive compatible. We randomly selected a price (an integer) between 1 and 100 to determine whether the participant receives the monetary reward or the meat bundle. Each price has the same chance of being extracted *independently* of the participant's choice in the first multiple price list. If the randomly selected price was not the one the participant had seen, we inferred his or her choice for this price from the choices for the other price levels. This strategy was feasible because we forced a single switching and hence we enforced consistency in choices.

At the end of the experiment, we asked the participants about their meat consumption patterns, attitudes toward climate change, and trust in the experimenters. We also asked for their contact information (both home address and email) to deliver the meat product or the monetary bonus, if any.

Implementation

We recruited participants on the platform Lucid between 31st March 2022 and 15th April 2022.[25] We focused on participants who consume meat and excluded those who lived outside contiguous US states due to shipment requirements by Porter Road.[26] 2,081 participants satisfied the pre-registered inclusion criteria: 1,047 were assigned to the NoInfo treatment and 1,034 were assigned to the Info treatment.[27] Participants are representative along gender and age. Table D.2.1 in the Online Appendix shows that demographic characteristics are balanced across treatments. Our sample is on average 46.8 years old (SD $= 17.1$) and 48.4% of the participants identified themselves as male. The median survey completion time was 17 minutes.

We implemented one of the two MPL decisions for one in every 20 participants and delivered the meat bundle (beef or poultry, depending on the selected MPL) or the monetary bonus, based on the participant's choice for the randomly selected price level. Finally, one (lucky) participant received a $500 completion reward. All bonus amounts were paid using Amazon gift cards. We preregistered our hypotheses and sample sizes on Aspredicted.org, the preregistration is available in the Appendix C.1.

4.4.1 Results

Following our preregistration, we focus on the belief and WTP data from the first part of the experiment for a clean analysis of the treatment effect.This means that belief and WTP data about the beef bundle come from BeefFirst treatments ($N = 1,048$) and the data about the poultry bundle come from PoultryFirst treatments ($N = 1,033$).

As in the climate survey discussed in Section 4.2, participants exhibited a significant underestimation of the size of CO_2 emissions from beef and poultry. Figure 4.7 shows that the magnitude and the prevalence of underestimation are more significant in the experiment as compared to the survey—median beliefs are much lower in the experiment (even though the quantity of meat products presented to the participants was more than twice as large as the quantity used in the survey) and the fraction of participants who underestimated the emission size was 92.7% for beef and 89.4% for poultry, respectively. Like in our survey, we see a large difference in the

[25]Lucid was acquired by Cint (`https://www.cint.com`) in January 2022, but still operated under the old name at the time of our experiment.

[26]To enhance data quality, we included five attention checks and three comprehension questions about the instructions. Participants were excluded if they failed any of the attention checks or if they needed more than five attempts to answer the comprehension questions correctly.

[27]Number of participants in each treatment is: 520 in the BeefFirst, Info treatment, 528 in the BeefFirst, NoInfo treatment, 514 in the PoultryFirst, Info treatment, 519 in the PoultryFirst, NoInfo treatment.

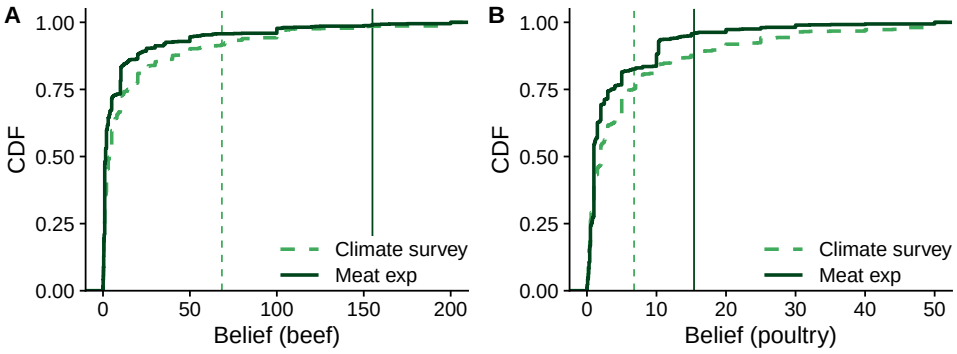


Figure 4.7: Empirical CDFs of beliefs about CO_2 emissions from two samples. (A) Beef. (B) Poultry. *Notes*: The size of meat products for belief elicitation was 7 oz in the climate survey and 1 lb (= 16 oz) in the meat experiment. For the data from the meat experiment, we focus on belief data from the first elicitation in the first part of the experiment. Vertical dashed lines correspond to the "true" size of CO_2 emissions (A: 155 miles for the meat experiment and 68.39 miles for the climate survey; B: 15.4 miles for the meat experiment and 6.78 miles for the climate survey).

size of underestimation between the two products: the absolute level of underestimation for the median subject is 153 miles for beef and 14.4 miles for poultry, respectively.

Participants were initially equally uninformed about CO_2 emissions across treatments. The distributions of prior beliefs (asked before WTP) show no differences between Info and NoInfo treatments for both meat products (Figure 4.8AB). Providing information successfully shifted the beliefs of many participants in the treated groups, as evident in jumps in the distributions of posterior beliefs (asked after WTP), illustrated in Figure 4.8CD. In particular, 64.8% $(337/520)$ of participants moved their beliefs to the correct value for beef, and 51.0% $(262/514)$ did so for poultry.[28]

Remember that our model in Section 4.3 predicts that information has a positive impact in the direction of reducing the demand for beef but has no impact on the valuation of poultry. In the experiment, these predictions are translated into a *decrease* in average WTP for the beef bundle and no effect for the poultry bundle. These predictions are not supported in the data. Figure 4.9A shows the WTP for meat products by treatment. If anything, there is a small *upward* movement in the valuation of the beef package after information provision. Average WTPs are not significantly different between treatments for both products (beef: $t(1046) = -1.200$, $p = 0.230$; poultry: $t(1031) = 0.938$, $p = 0.349$). Panels B and C of Figure 4.9 give a more complete overview of demand and show the proportion of buyers for each price, confirming that there is no discernible difference between the treatments.

Table 4.3, column (1) shows the effect of information on beef valuation in a regression analysis. This "null" finding is robust to the inclusion of several control variables in the regression (Table D.2.2 and Figure D.2.6 in the Online Appendix). Several of those covariates have sensible signs: we find a higher WTP for beef for those subjects who report above-average beef consumption, or who report that it is difficult to reduce beef consumption. We also find a lower WTP for both beef and poultry amongst women and younger individuals. Finally, in Online

[28] If we allow a margin of $\pm 10\%$, the number increases to 68.3% $(351/514)$ for poultry.

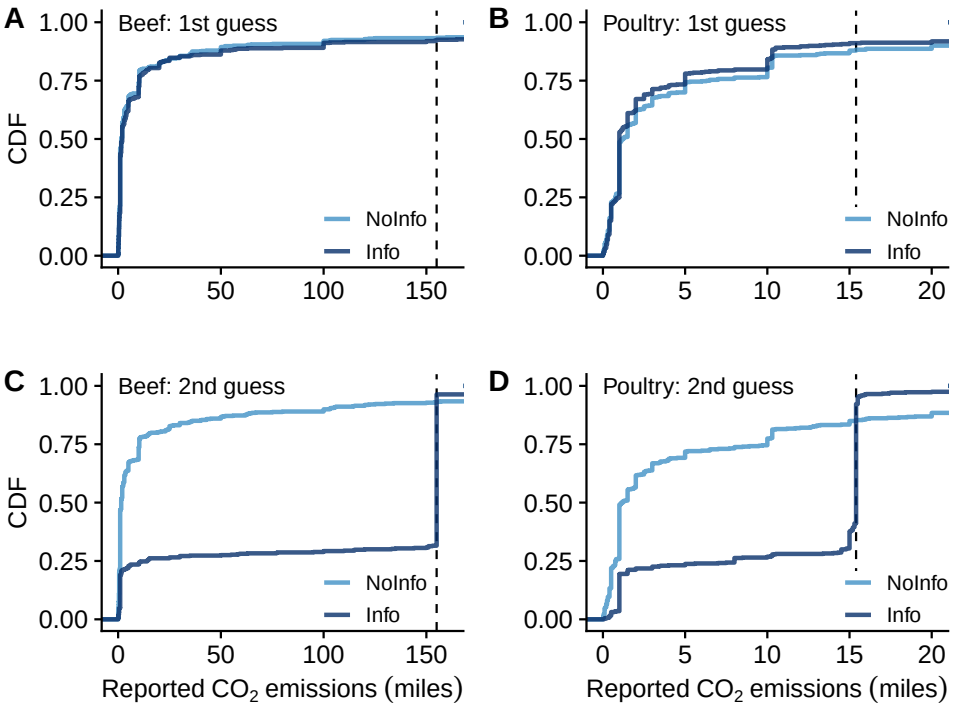


Figure 4.8: Beliefs about CO_2 emissions from two meat products. *Notes*: We focus on the data from the first part of the experiment (panels AC: BeefFirst treatments; panels BD: PoultryFirst treatments). Vertical lines correspond to the "true" size of CO_2 emissions (15.4 miles for poultry and 155 miles for beef).

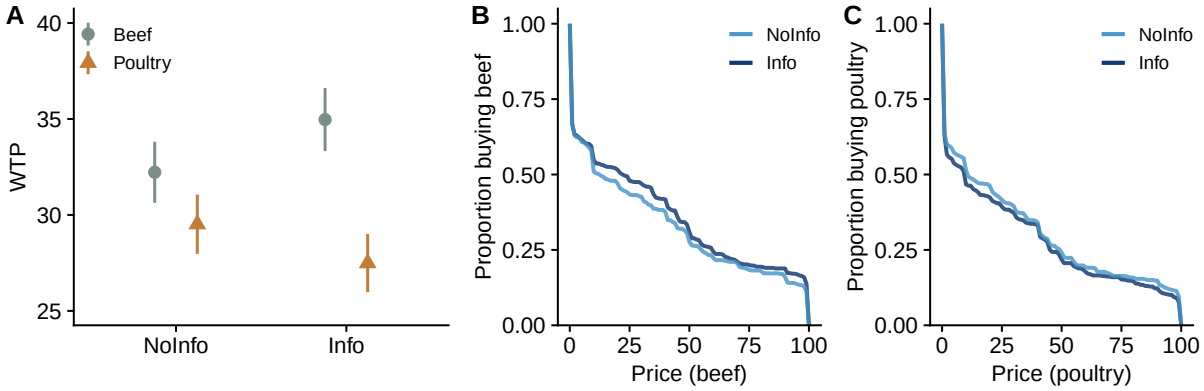


Figure 4.9: (A) Average willingness to pay for the first meat product. (BC) The proportion of participants buying the meat product at each price. *Notes*: We focus on the data from the first part of the experiment. In panel A, Bars indicate SEM. Figure D.2.4 in the Online Appendix shows the CDFs of WTPs.

Appendix Figure D.2.6 we also conduct an analysis of the treatment effect by subgroup. For all subgroups, we cannot reject the null hypothesis that the information effect for beef is zero.

4.4.2 Interpretation of the Null Effect

We now turn to investigate possible reasons for the observed null effect of information about CO_2 emissions on the demand for meat. We focus on beef, where we predicted that information should affect willingness to pay negatively and decisively.

Were participants' beliefs insensitive to the information treatment? In both the Info and the NoInfo treatments, we measure beliefs twice (Figure 4.6). In the Info treatment, the second belief, or posterior, is measured after information about beef consumption is provided. Participant's posterior is affected by the information treatment and exhibits, on average, less optimism about CO_2 emissions (see Figure 4.8CD). This shows that participants' beliefs were changed by the information they saw. However, these belief changes do not translate into differences in WTP. Column (2) of Table 4.3 shows regression results of WTP on a dummy for the Info treatment, including only participants in the latter treatment who responded to information by updating their beliefs upward. While the coefficient on the Info treatment declines relative to the full sample (column (1)), the null effect remains.

Did participants become more pessimistic about other meat products? It is possible that information about beef made participants more pessimistic about other meat products. This would limit the options for (low carbon) substitution, rendering demand for beef inelastic in information. We can address this point in several ways. The first is to directly control for this spillover in beliefs. In the BeefFirst treatment, we measure participants' beliefs about the CO_2 emissions associated with poultry after the participants received information about and stated their willingness to pay for beef. We find that participants do indeed become much more pessimistic about poultry after receiving information about beef. About 63% of the participants in the BeefFirst, Info treatment ($317/505$) overestimated the size of CO_2 emissions from poultry (reported numbers above 15.4 miles) and 48 subjects reported 155 miles, which is exactly the size of CO_2 emissions from beef they learned about in the first part of the experiment (see Figure D.2.5 in the Online Appendix). However, this updating about a substitute product does not appear to be an important mediator of the information effect on beef demand: the null effect persists after controlling for the beliefs associated with poultry consumption (Table 4.3, column (3)).

In addition, We can look at the case where beef is the second product participants can buy. Here, by the time participants state their willingness to pay for beef in the Info treatment, they have received information on both poultry and beef. This group is therefore aware of a climate-friendly substitute. However, we find no treatment effect in the second product either (Table 4.3, column (8)).

We can also test for this confound using participants' stated intentions about future consumption in the post-experimental questionnaire. At the end of the experiment, we asked participants "Do you intend to reduce your beef/poultry consumption in light of its CO_2 emissions?" and they answered on a Likert scale from 1 to 5. In the Info treatment, participants will have received information about both beef and poultry by the time they answer this question, and hence know about poultry being a low-carbon substitute for beef. Yet, a chi-squared test of independence shows no differences in response distribution between Info and NoInfo treatments for intention to reduce beef (Figure 4.10; $\chi^2(4) = 0.964$, $p = 0.915$).

Preaching to the choir: Is there a mismatch between who is optimistic about the CO_2 emissions associated with meat consumption and who cares about mitigating CO_2 emissions?

Table 4.3: Interpretation of the null effect of information on WTP for beef.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Info	2.743	0.528	2.914	1.673	5.277		4.835	−0.860
	(2.285)	(2.448)	(2.400)	(2.847)	(3.284)		(3.191)	(2.292)
Belief (poultry)			−0.011					
			(0.012)					
Difficulty (beef)						3.074***		
						(1.004)		
Constant	32.225***	32.225***	33.018***	32.912***	34.574***	25.580***	35.062***	33.080***
	(1.590)	(1.590)	(1.660)	(1.984)	(2.263)	(2.942)	(2.162)	(1.616)
First product	Beef	Beef	Beef	Beef	Beef	Beef	Beef	Poultry
Observations	1,048	901	1,032	672	529	991	579	1,013
R^2	0.001	0.0001	0.002	0.001	0.005	0.010	0.004	0.0001

Notes: The dependent variable is WTP for beef. Samples are as follows. (1) All participants in the BeefFirst treatments. (2) Participants in the NoInfo treatment, and those in the Info treatment who responded to information by updating their beliefs upward. (3) Participants who completed both parts of the experiment. One subject who reported an extremely large belief ($\approx 1.46 \times 10^9$) about CO_2 emissions from poultry is excluded. (4) Participants in the BeefFirst treatments who self-proclaimed to care about the environment (based on the response to the question "How severe do you consider the problem of climate change?"). (5) Participants in the BeefFirst treatments who self-reported consuming beef at least three times per week. (6) Participants in the BeefFirst treatments who self-reported consuming beef. (7) Participants in the BeefFirst treatments who expressed trust in us actually sending meat. (8) All participants in the PoultryFirst treatments. Robust standard errors are reported in parentheses. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

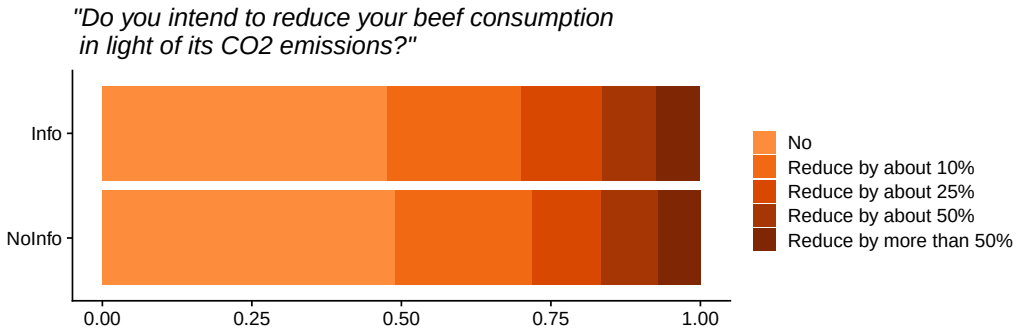


Figure 4.10: Distribution of responses to a survey question: "Do you intend to reduce your beef consumption in light of its CO_2 emissions?" *Notes*: We focus on 1,013 participants in the BeefFirst treatments. Participants responded on a 5-point Likert scale. (1: "No." 2: "Yes, I am prepared to reduce my current consumption by about 10%." ... 5: "Yes, I am prepared to reduce my current consumption by more than 50%.")

One reason that information may have little impact on CO_2 emissions is that prior optimism about CO_2 emissions is concentrated among individuals who have little willingness to mitigate. The info treatment would then correct the beliefs of only those who have no interest in mitigation. Our structural model was explicitly designed to make predictions that take this mismatch into account, so our initial predictions, based on the representative survey, are not subject to this concern.

To see whether these concerns could matter in the second dataset, we can restrict our analysis to those participants who self-proclaim to care about the environment. The null effect persists in this restricted sample (Table 4.3, column (4)).

Are near-vegetarian driving the results? If only near-vegetarians are optimistic about the CO_2 emissions associated with meat consumption, then providing this information will do little to curb the demand for meat. Of course, this state of affairs is ex-ante implausible, but, for the sake of completeness, we can provide an explicit test for this hypothesis by restricting our dataset to participants who consume meat at least three times per week (i.e., above-median frequency). The null effect persists in this restricted sample (Table 4.3, column (5)).

Do participants suffer from an intention-action gap? An intention-action gap would manifest itself as a stated intention to reduce meat consumption in the future, but a failure to do so in the present. The underlying reason could be a preference for immediate gratification or a self-control problem. Yet, as we reported above, and as Figure 4.10 shows, intentions to reduce beef consumption are not much affected by the information treatment. Thus, the null effect of information does not stem from a failure to implement virtuous plans, but from a failure to make such plans, to begin with.

Does the information cause participants to decrease their consumption of lower-quality meat outside of the experiment? A key challenge of our experimental setup is to sell participants a product that they find appealing. To this end, we used high-quality meat. But this may invite

the concern that participants respond to information by demanding less, but better, meat. If this were the case, then the information treatment may decrease average meat consumption, but not the willingness to pay for the meat we sell to participants. Again, the fact that information does not impact participants' stated intention to consume beef rules out this conjecture.

Do participants react to information not by demanding less beef, but by offsetting the CO_2 emissions of their consumption outside of the experiment? We deem this hypothesis unlikely. It requires individuals to care about mitigating CO_2, to take into account and feel the pain of their meat consumption emitting CO_2, but to be completely inelastic in their meat consumption. Empirically the price elasticity of demand for beef steaks in the US is between -0.42 and -0.52, making beef demand far from inelastic (Dong, Davis and Stewart, 2015). So if learning about the CO_2 emissions increases the subjective cost of buying meat, it seems unlikely that participants do not use the rather elastic margin of adjustment that is a decrease in the WTP for meat, and instead adjust only buy purchasing offsets outside of the experiment.

Does our willingness to pay measure suffer from noise, misinterpretation, or lack of trust? A possible reason for a null effect of the information treatment may be that our measure of demand is very noisy. If our WTP measure is a very poor proxy for actual demand, then it would follow that this measure does not necessarily change with new information, even if this information would have had an impact on participants' actual demand for meat. To shed some light on this possible reason for a null effect, we ask whether our willingness to pay measure is correlated with other measures of preferences for meat. This would not be the case if WTP was very noisily measured. We find that WTP for beef is significantly correlated with participants' self-reported difficulty in reducing beef consumption if they had to (Table 4.3, column (6)).

A related worry may be that despite our elaborate efforts to be credible, (some) participants did not believe us that we would actually send them the meat they purchased with positive probability. Then, what they answered in the willingness to pay elicitation may not reflect their sincere demand for beef. To test this hypothesis we ask whether there was a treatment effect among those who expressed a lot of trust in us actually sending meat in the post-experimental survey.[29] The null effect persists in this restricted sample (Table 4.3, column (7)).

A final, somewhat related, concern is that the participants misunderstood our WTP question and thought they had to indicate the (socially) fair price for the beef shipment. This misunderstanding could generate a null result if some participants in the Info treatment thought that the fair price should be higher due to the high emissions.

Several considerations assure that this misunderstanding is unlikely. First, the word "price" did not appear in the experiment: subjects made a sequence of binary buying decisions from which we infer a WTP. Second, we advised the participants to use their valuation of the meat to make their decisions. Third, the instructions did not contain any reference to CO_2 offsets or to other environmental actions associated with the product (and indeed there was no such

[29]Participants responded to the question "Do you trust that the researchers will indeed ship meat products as described in the instructions?" on a 5-point Likert scale (1: not at all; 5: completely).

offset), so there is no reason to pay more out of fairness concerns. Finally, if the information made participants think that the fair price is higher, we should find that information reduces the intention to consume beef. However, as we discussed above, we do not find evidence for this treatment effect.

Was the null effect a fluke? Even relatively well-powered studies may sometimes result in erroneous null effects. Three results speak against this hypothesis. First, we can ask whether there is any correlational evidence that beliefs about CO_2 are predictive of the willingness to pay for meat. While any such evidence is subject to the usual caveats and endogeneity concerns, a strong negative correlation between beliefs about CO_2 emissions and WTP in the NoInfo treatment should give us pause in interpreting the null effect of the info treatment. We find that prior beliefs in the NoInfo treatment do not correlate with meat consumption.

Second, we can use the comparison of the Info and NoInfo treatments when beef was offered in the second part as a replication experiment. Of course, because these data stem from Part 2 of the experiment, the treatment comparison is less tightly controlled, with information about poultry possibly also bearing on participants' willingness to pay for beef. At the same time, it is hard to construct an explanation of how this additional information would lead to a null effect. We find that experiment 2 also features null effects of the information treatment.

Third, the lower bound of the 95% confidence interval for the effect of information on the willingness to pay for beef is $-\$1.74$. Hence, even if the information has an effect that we are not powered to detect, this effect is likely less than 2% of the market price of the meat.

Finally, and as we have already seen, the information does not affect participants' stated intention to reduce meat consumption.

What, then, causes the null effect? Having ruled out several possible explanations for the observed null effect, we are led to conclude that people's decision to eat meat appears not to be subject to concerns about associated CO_2 emissions. That is, even though we see that people are willing to invest in emission reduction when this willingness is elicited directly, their desire to curb emissions in meat consumption appears to be drowned out by the many other considerations that go into their consumption decision. If this is the reason behind the null effect, then we should be no more optimistic about finding an effect of information in still "wilder" settings. After all, we made sure that our information actually moved beliefs and we can be confident the climate impact of various consumption activities was a salient feature of the decision making environment.

4.5 Conclusions

We have used incentivized survey techniques to elicit both beliefs about the carbon impact of consumer products and the valuation of this impact. We find that most consumers underestimate the impact, but heterogeneity is large. While they are willing to pay to offset carbon emissions,

this willingness is highly concave and varies by subgroups. We use these inputs in a simple structural model to predict the impact of information. In an experimental test, we find little support for our predictions: despite a large correction in their beliefs about beef meat, subjects are largely unresponsive in their valuations of beef products.

Our results show that correcting consumer beliefs does not necessarily lead to lower demand for carbon-intense consumer products, even in settings where misperceptions are large, and consumers indicate that they are interested in offsetting emissions. This suggests that the climate impact of behavior is not a strong motivating force for most consumers in our experiment. Our findings are not inconsistent with those of experiments that find that the effects of carbon labels are small and short-lived. Because our design keeps the salience of climate change constant across conditions, we show that pure shifts in beliefs do little to change consumption behavior. This suggests that results in these other experiments are driven at least in part by increasing the salience of the climate change phenomenon, or by highlighting the emerging social norms around low-carbon consumption.

Our results also speak to the implications that can and cannot be drawn from some existing evidence. Evidence of widespread misperception of the climate impact of different consumption behaviors has sometimes been used to argue that information campaigns can lead to meaningful change. We show that this is not necessarily the case. Similarly, other papers have investigated attitudes toward climate change by using donation decisions, willingness to mitigate and survey responses. The results from these papers may be important in their own right, but our results temper confidence that these measures translate directly into everyday behavior like food consumption.

In fact, the picture that emerges from our and other studies is that the immediate return on information provision policies does not justify their current popularity among policy makers. It suggests that relying on the good intentions of informed individuals will not by itself deliver the important changes that we need in our carbon consumption, and that we may need to rely on more systemic approaches (Chater and Loewenstein, 2022). Of course, our results leave open the possibility that other types of information provision, in a different context or evaluated using a different metric will be more effective in changing behavior. Having more informed citizens may also have other beneficial effects through long-run reflective processes, for instance by increasing political support for a carbon or meat tax. Future research should help elucidate such mechanisms.

Appendices

Appendix A

Fair Shares and Selective Attention - Appendix

A.1 Preregistration

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Tracking fairness (#44417)

Created: 07/12/2020 01:54 AM (PT)
Shared: 11/03/2020 12:17 PM (PT)

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?
We study the origin of self-serving biases in monetary allocation problems. If people are randomly placed in a (dis)advantaged position, how does this affect their attention to meritocratic information, the ethical criteria for making decisions, and the subsequent allocation choices? Detailed hypotheses are specified in point 5).

3) Describe the key dependent variable(s) specifying how they will be measured.
In Part 1 of the experiment, subjects first produce a surplus together with a matched partner on several tasks. We create variation in contribution to the surplus by randomly giving one of the partners a higher piece rate than the other. In Part 2 of the experiment, some subjects are given information on the performance on the tasks as well as the total contribution, and make allocation decisions in the role of dictator. We use Mouselab to track the way subjects explore information about task performance.

Per every decision of the dictator we record:
- the split in the total surplus between dictator and recipient.
- dwelling time (mousetracked) on each of the following information 1) the dictator & recipient contribution to the pie in monetary terms, 2) the number of answers in the task the dictator & recipient got correct.

4) How many and which conditions will participants be assigned to?
Subjects are assigned to be “receivers” and “dictators”. Both groups take part in a series of performance tasks to determine the surplus. We are mostly interested in the dictators.

All dictators are assigned to one of two treatments:
Advantaged: receives a high piece rate per correct answer in the task.
Disadvantaged: receives a low piece rate per correct answer in the task.

Each dictator participates (in this order) in an
Involved condition: 20 allocations between themselves and another randomly matched participant
Benevolent condition: 20 allocations between two other participants.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Hypothesis 1 (Behavior): In the involved condition, advantaged dictators give less money to the receivers than disadvantaged dictators.

We test this hypothesis with a non-parametric rank sum test. We will perform regressions to control for subject characteristics with standard errors clustered for each participant.

Hypothesis 2 (Attention): In the involved condition, advantaged dictators spend relatively less time on correct answer information and more time on monetary contribution information than disadvantaged dictators.
Across dictator groups, we investigate total time looking at information as well the proportion of time spent looking at correct answers, using a non-parametric rank sum test. We will also perform regressions with standard errors clustered for each participant.

Hypothesis 3 (Persistence): The effects documented in 1) and 2) persist in the benevolent condition.
The tests are the same as for Hypothesis 1 and 2, but now in the benevolent condition. We will also compare the effects in both conditions using a difference in difference approach.

Hypothesis 4 (Role of attention): Attention patterns drive giving decisions.
For correlational evidence, we use regressions to investigate how sensitive the treatment effect (Hypothesis 1) is to controlling for total and relative looking time. For a causal inference, we use an instrumental variable analysis to exploit variation generated by the (randomly varied) orientation of patterns on the

screen.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Following standard Mouselab protocols, we will exclude information that was revealed for less than 200 ms.

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.
We will recruit 200 dictators from the online platform Prolific. These are divided 50-50 between the advantaged and disadvantaged condition. We recruit the corresponding number of recipients.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will conduct a number of secondary analyses:
- We will compare by treatment the fairness criteria people list in the questionnaire as being most socially appropriate.
- We compare by treatment the fairness "types" based on Cappelen et al. (2007), and correlate these types with attentional patterns.
- Correlate attention, behavior and political preferences elicited in the final questionnaire.
In addition, we will explore additional measures of attention, and their explanatory power for giving decisions. We will conduct robustness analysis on the revelation threshold in point 6).

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Tracking fairness - attention manipulation (#52512)

Created: 11/18/2020 09:42 AM (PT)
Shared: 02/16/2021 06:11 AM (PT)

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?
We study the origin of self-serving biases in monetary allocation problems. If people are randomly placed in a (dis)advantaged position, how does this affect their attention to meritocratic information, the ethical criteria for making decisions, and the subsequent allocation choices? In a previous version of the experiment, we showed that advantaged dictators pay less attention to information that reveals pure merit (actual task performance). In this experiment we ask how randomly induced variations in attention affect decision making.

3) Describe the key dependent variable(s) specifying how they will be measured.
In Part 1 of the experiment, subjects first produce a surplus together with a matched partner on several tasks. We create variation in contribution to the surplus by randomly giving one of the partners a higher piece rate than the other. In Part 2 of the experiment, some subjects are given information on the performance on the tasks as well as the total contribution, and make allocation decisions in the role of dictator. We manipulate how long different kinds of information are available to people.

Per every decision of the dictator we record:
- the split in the total surplus between dictator and recipient.
- dwelling time (mousetracked) on each of the following information 1) the dictator & recipient contribution to the pie in monetary terms, 2) the number of answers in the task the dictator & recipient got correct.

4) How many and which conditions will participants be assigned to?
Subjects are assigned to be “receivers” and “dictators”. Both groups take part in a series of performance tasks to determine the surplus. We are mostly interested in the dictators.

All dictators are assigned to one of two treatments:
Advantaged: receives a high piece rate per correct answer in the task.
Disadvantaged: receives a low piece rate per correct answer in the task.

We cross-randomize these treatments with another dimension:
Merit focus: in a majority of trials, the information about task performance (merit) is available longer.
Output focus: in a majority of trials, information about total contribution to surplus is available longer.

Each dictator participates (in this order) in an
Involved condition: 20 allocations between themselves and another randomly matched participant
Benevolent condition: 20 allocations between two other participants

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We test two main hypotheses for both the involved and the benevolent dictators:
1) Dictators in the “Merit Focus” treatment will give more to disadvantaged recipients.

We will test this in a regression with data for all trials and a dummy for all trials with Merit Focus, as well as controls for subject and trial characteristics.

2) Compared to a situation with freely chosen attention, making dictators look longer at “inconvenient” information (i.e. ”Merit focus” for advantaged dictators, “Output focus” for disadvantaged dictators) will reduce the relative bias of advantaged dictators towards the advantaged recipients.

We combine the data from this experiment with a previous experiment in which dictators could freely choose what to look at. We will use regressions to evaluate the “difference in difference”.

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

Following standard Mouselab protocols, we will exclude information that was revealed for less than 200 ms.

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.
We will recruit 400 dictators from the Prolific platform. Dictators will be evenly split between the 4 between subject conditions (i.e. 100 in each cell). We recruit a corresponding number of receivers.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will investigate whether the impact of merit/output information on giving differs between advantaged and disadvantaged dictators.
We will correlate giving and attention with several additional elicitations in the questionnaire on perceptions of fairness.

A.2 Model

This section develops a model that investigates the relationship between attention and redistribution decisions and between selective attention and self-serving biases. The proofs of the results are provided Appendix A.3.

A.2.1 Set-up

Imagine two people, a dictator (she) and a recipient (he), indicated with subscript $i = 1$ and $i = 2$ respectively. In the production phase, agents produce w_i leading to joint endowment $W = w_1 + w_2$. We model the role of effort and luck multiplicatively, as in the experiment: $w_i = e_i l_i$. Here, $e_i \in [\underline{e}; \bar{e}]$ is the effort exerted by i, where $\underline{e} > 0$ is the minimum possible effort and $\bar{e} > \underline{e}$ is the highest possible effort. The luck component is the multiplier on effort $l_i \in L, H$ with $0 < L < H$. The dictator and the recipient differ in their luck, i.e. $l_1 \neq l_2$. We call the dictators for which $l_1 = H$ "Advantaged" and the dictators for which $l_1 = L$ "Disadvantaged". After the production phase, the dictator decides the allocation of the endowment (x_1, x_2) such that $x_1 + x_2 = W$.

The dictator knows W and whether she is Advantaged or Disadvantaged, that is, whether $l_1 > l_2$ or vice versa. However, she does not know the exact values of l_1 and l_2, but she knows they have a distribution $f_{l_1}(l_1)$ and $f_{l_2}(l_2)$. Moreover she believes that $e_i \sim f_e(e_i)$, a symmetric distribution. The effort levels of the two agents are independent and identically distributed. To resolve the uncertainty about e_i and l_i the dictator can access a signal about (e_1, e_2) and another signal about (w_1, w_2). These signals resolve uncertainty fully and are available for free, although paying attention to them may be costly, as we discuss below.[1]

Dictator preferences. The dictator's preferences are represented by

$$U(x_1, \mathbf{t}) = u(x_1) - g(x_1 - r(\boldsymbol{t})) - C(\mathbf{t}, \bar{\mathbf{t}}). \tag{A.1}$$

Here, $u(x_1)$ is the utility from her monetary allocation, which is increasing and concave. The second component captures guilt from an unfair allocation, which may depend on attention. The last component is an attentional cost. We now discuss these final components in detail.

Fairness. The term $g(x_1 - r(\boldsymbol{t}))$ indicates the guilt cost the dictator pays if she keeps more than $r \in [0; w]$, the amount the dictator considers to be fair to keep for herself. We assume that $g : [0; w] \to \mathbb{R}_+$, $g(x_1 - r) = 0$ if $x_1 \leq r$, and $g(x_1 - r) > 0$ if $x_1 > r$. $g(x_1 - r)$ is twice differentiable, increasing and strictly convex if $x_1 > r$. Modeling fairness concerns as disutility (guilt) from the difference between the actual and the fair share is common in the literature (Konow, 2000; Rodriguez-Lara and Moreno-Garrido, 2012; Cappelen et al., 2007, 2013).

[1] This setup follows our experimental design. Our model also works if dictators are perfectly informed, and attention to different kinds of information only serves to contemplate the associated fairness criterion.

We assume the fair amount depends on a weighted sum of three different fairness criteria:

$$r(\mathbf{t}) = \pi_\mu(\mathbf{t})x_\mu + \pi_\lambda(\mathbf{t})x_\lambda + \pi_\eta(\mathbf{t})x_\eta.$$

Here, x_k is fair amount the dictator should keep according to criterion $k \in \{\mu, \lambda, \eta\}$:

- The **meritocratic criterion** $x_\mu := W\frac{e_1}{e_1+e_2}$ prescribes keeping an amount proportional to the dictators effort.
- The **libertarion criterion** $x_\lambda := w_1$ prescribes keeping an amount proportional to the dictators' output, without correcting for luck.
- The **egalitarian criterion** $x_\eta := \frac{W}{2}$ prescribes keeping half of the output.

The weight $\pi_k(\mathbf{t})$ depends on attention vector $\boldsymbol{t} = \{t_\mu, t_\lambda, t_\eta\}$, where t_k indicates the timespan that dictator attends to criterion k (see more details below). We assume that $\frac{\partial \pi_k}{\partial t_k} > 0$, so weights increase in the attention paid to the corresponding criterion. Since we normalize vector $\boldsymbol{\pi}$ is such that $\sum_{k\in K} \pi_k(\mathbf{t}) = 1 \; \forall \, \boldsymbol{t}$, this implies that $\frac{\partial \pi_k}{\partial t_{-k}} < 0$.

This way of modelling distortions due to attention is adapted from Bordalo, Gennaioli and Shleifer (2021). The positive relation between attention and decision weights is supported by Pärnamets et al. (2015), who exogenously manipulates the time participants spend looking at two statements regarding controversial moral topics. They find that participants are more likely to endorse the statement that they look at longer. Ghaffari and Fiedler (2018) replicated and extended this finding. Section 1.5.2 discusses how visual attention changes what the dictators consider a fair allocation in our experiment.

Attention. The attention vector $\boldsymbol{t} = \{t_\mu, t_\lambda, t_\eta\}$ captures two types of attention. First, it captures visual inspection of information. Thus, we will denote by t_μ the time spent accessing information about efforts e_i, as this is relevant exclusively for the meritocratic criterion. Similarly, we denote by t_λ the time spent attending to outputs w_i, as this is relevant exclusively for the libertarian criterion. Second, visual information may be accompanied with various types of information processing and introspective contemplation to evaluate the proper use of the criterion. Indeed, the well-established eye-mind theory shows that visual attention is accompanied by the processing of the underlying information (Just and Carpenter, 1980). We assume all these aspects are captured by t. This also allows us to model attention egalitarian split t_η, which does not require visual attention to any of the production data, but is relevant for Proposition 3 below. Proposition 1 and 2 go through when we drop attention to the egalitarian criterion and we assume that there is a fixed weight the dictator gives to the egalitarian criterion.

We make the following assumptions about t.

1. **Attention budget.** The dictator has a total time of T to attend to information, and needs to spend all this time looking at information such that $t_\mu + t_\lambda + t_\eta = T$.
2. **Top-down control.** The dictator has control over her attention. That is, she chooses a vector of attention $\boldsymbol{t} \in S$, where S is a 2-simplex of edge length T.

3. **Bottom-up salience.** Attentional control is costly, as certain states may be salient and naturally attract attention. The importance of salience for decision making are documented in an well established literature in psychology and a growing literature in economics Bordalo, Gennaioli and Shleifer (2021). To capture this, we assume there exists a *default* bottom-up attention pattern $\bar{\boldsymbol{t}} = \{\bar{t}_\mu, \bar{t}_\lambda, \bar{t}_\eta\}$ with $\bar{t}_\lambda > 0$, $\bar{t}_\mu > 0$, and $\bar{t}_\eta \geq 0$.[2] When the dictator deviates from default $\bar{\boldsymbol{t}}$, she pays a cost $C(\boldsymbol{t}, \bar{\boldsymbol{t}}) : S \to \mathbb{R}$, which is twice differentiable and increasing in $|t_k - \bar{t}_k| \; \forall \; k \in K$. This cost captures both the concentration costs of manipulating attention and the psychological cost of diverting attention in order to self-deceive about the size of r and hence increase x_1. We normalize $C(\bar{\boldsymbol{t}}, \bar{\boldsymbol{t}}) = 0$.

4. **Curiosity.** $C(\boldsymbol{t}, \bar{\boldsymbol{t}}) > u(W)$ if $t_\lambda = 0$ or if $t_\mu = 0$. This assumption models curiosity, as it assures ignorance is too costly for the dictator. Golman et al. (2021) shows how curiosity is an important driver for information acquisition, in particular when it is salient that information is available. In our data, most dictators access all information (see Section 1.5.5 for more details).

5. **Speed of learning.** Minimal attention is needed to acquire the information about (w_1, w_2), and (e_1, e_2) and resolve all the uncertainty. If $t_\lambda > 0$, the dictator knows the exact values of (w_1, w_2). If $t_\mu > 0$, the dictator knows the exact value of (e_1, e_2). This assumption reflects the fact that the information is very simple. Four numbers are all that the dictators have to learn. In the experiment, these numbers are mostly one or two digits.

Timeline. The timeline is as follows:

$\tau = 0$ Production task.

$\tau = 1$ The dictator receives perfect information on W and on whether she is Advantaged or Disadvantaged. Furthermore, she can allocate her attention, and access information about e_1, e_2, w_1, and w_2. The time she spends on the different types of information determines **t**.

$\tau = 2$ The dictator splits W in x_1 and x_2.

The dictator maximizes her utility by choosing **t** and x_1 sequentially. To solve the model, we therefore work backwards, first computing the optimal choice for a given level of attention, and then maximizing the level of attention given the resulting choice.

[2]Note that default $\bar{\boldsymbol{t}}$ is a function of the information. That is: $\bar{\boldsymbol{t}}(e_1, e_2, w_1, w_2) : \mathbb{R}^4 \to S$. Note that $\bar{\boldsymbol{t}}$ depends only on the decision making environment, and not on the dictators characteristics (e.g. whether she is Advantaged or Disadvantaged). That is $\bar{\boldsymbol{t}}(e_1, e_2, e_1H, e_2L) = \bar{\boldsymbol{t}}(e_1, e_2, e_1L, e_2H)$. The dependence of the bottom-up vector of attention on the information highlights that the bottom-up process influence the dictator only if she access the information. We don't formally model what happens when the dictator avoids all or part of the information because Lemma 1 shows that the dictator always access the information about (e_1, e_2, w_1, w_2).

A.2.2 Results.

We first show that our model predicts selective attention and self-serving biases in allocation decisions. We then demonstrate that Advantaged dictators keep a larger amount for themselves than the Disadvantaged ones. Finally, we turn to the impact of implementing exogenous restrictions on attention, as in our constrained focus treatments.

Selective attention. Let's call t_μ^{*A} and t_λ^{*A} the optimal level of attention to information about merit and about outcome if the dictator is Advantaged and t_μ^{*D} and t_λ^{*D} the optimal level of attention if she is Disadvantaged. We can define Δ Attention$^A = t_\mu^{*A} - t_\lambda^{*A}$ and Δ Attention$^D = t_\mu^{*D} - t_\lambda^{*D}$.

Proposition 1 (Selective Attention)**.** *Δ Attention^A $<$ Δ Attention^D. That is, compared to Disadvantaged dictators, Advantaged ones spend relatively less time looking at information about effort and relatively more time looking at information about outcome.*

Intuitively, dictators distort their attention to believe that they deserve a larger share of the endowment and hence reduce their guilt over keeping a larger share. Advantaged and Disadvantaged dictators, however, distort attention in opposite directions. The Advantaged dictators move attention from merit information to outcome information because they receive more if they implement a libertarian rather than a meritocratic split. The opposite is true for the Disadvantaged dictators: they shift their attention from the outcome to the merit information because they receive more from a meritocratic rather than from a libertarian split.

Corollary 1.1 (Attention as a mediator of self-serving biases)**.** *Selective attention allows dictators to act more selfishly.*

This result follows immediately from Proposition 1. Selective attention reduces the marginal guilt cost for any amount the dictator keeps for herself. As a result, the dictator keeps more.

Restricting attention. Our objective in this paragraph is to check whether the model predicts that Advantaged agents receive more money in the Outcome Focus treatment than in the Merit Focus treatment. To do so we need to formalize our two attention manipulations in our experiment, let's call them *Mer* and *Lib*. Without loss of generality, we assume that *Mer* is the manipulation that restricts attention to (e_1, e_2), while *Lib* restricts attention to (w_1, w_2). Hence *Mer* models the Outcome Focus treatment, and *Lib* models the Merit Focus treatment. The manipulations restrict the set of vectors of attention among which the dictators can choose. Let's call $\hat{S}^{Mer} \subset S$ and $\hat{S}^{Lib} \subset S$ the two sets of feasible vectors of attention when the manipulations are in place. Moreover, let's define t_μ^{*Mer} and t_λ^{*Mer} the optimal attention to the meritocratic and libertarian criteria in $\hat{S}^{Mer}$. Similarly, define t_μ^{*Lib} and t_λ^{*Lib} the optimal attention to the meritocratic and libertarian criteria in $\hat{S}^{Lib}$. Finally, define ΔAttention$^{Mer} = t_\mu^{*Mer} - t_\lambda^{*Mer}$ and ΔAttention$^{Lib} = t_\mu^{*Lib} - t_\lambda^{*Lib}$.

Armed with the definitions above, we can go back to our experiment and study three properties of the attention manipulations. First, from Table 3.1, we can see that ΔAttention^{Lib} $>$

$\Delta\text{Attention}^{Mer}$. Second, from the same table, we see that the dictators spend a similar amount of time looking at Merit and Outcome information in the Merit Focus and in the Outcome Focus treatments. We can approximate this finding assuming that $t_\mu^{*Mer} + t_\lambda^{*Mer} = t_\mu^{*Lib} + t_\lambda^{*Lib}$. Finally, from the design we derive that the total time people spend thinking about the different criteria is the same independently of the treatment: subjects have 6 seconds on the information screen and then they are automatically redirected to the decision screen. Hence, $\sum_{k\in K} t_k^{*Mer} = \sum_{k\in K} t_k^{*Lib}$.

The proposition below shows that these three properties of the attention manipulations are sufficient conditions for Advantaged agents to receive more money under attention manipulation *Mer* than under attention manipulation *Lib*.

Proposition 2 (Effect of constrained Attention on Allocation)**.** *If* $\Delta Attention^{Lib} > \Delta Attention^{Mer}$, $t_\mu^{*Lib} + t_\lambda^{*Lib} = t_\mu^{*Mer} + t_\lambda^{*Mer}$, *and* $\sum_{k\in K} t_k^{*Lib} = \sum_{k\in K} t_k^{*Mer}$, *then the Advantaged agents receives more money if* $\boldsymbol{t} \in \hat{S}^{Mer}$ *than if* $\boldsymbol{t} \in \hat{S}^{Lib}$.

The result obtains because a lower ΔAttention decreases the weight the dictator gives to the meritocratic criterion and increases the weight she gives to the libertarian one. As a result, Advantaged dictators keep more money and Disadvantaged ones keep less.

While Proposition 2 predicts our main empirical result, a puzzling finding from our experiment is that the Advantaged dictators react more to our attention manipulations than Disadvantaged dictators. The proposition below shows that the model predicts this finding under some reasonable simplifying assumptions about a) the functional form of the utility function b) the attention process c) the characteristics of the attention manipulations. These assumptions are sufficient but not necessary for deriving the result.

Assumption 1 (Simplifying assumptions about the utility function)**.**

$$
\begin{aligned}
&u(x_1) = x_1 \\
&g(x_1 - r) = \frac{1}{2}\beta g(x_1 - r)^2 \\
&C(\boldsymbol{t}, \bar{\boldsymbol{t}}) = \frac{1}{2}\gamma[(t_\lambda - \bar{t_\lambda})^2 + (t_\mu - \bar{t_\mu})^2 + (t_\eta - \bar{t_\eta})^2]
\end{aligned}
$$

The first line of the assumption states utility function is linear in money. This is a good approximation for small stakes like the ones in our experiment (Rabin, 2000). The second line says that the guilt function is quadratic. This is a common assumption in the literature on fairness norms (Cappelen et al., 2007; Bortolotti et al., 2017). The third line states that the cost of attention distortion is a sum of quadratic costs. We chose this quadratic form for consistency with the functional form of the guilt function.

Assumption 2 (Simplifying assumptions about the attention process)**.**

$$r(\boldsymbol{t}) = \frac{t_\lambda}{t_\lambda + t_\mu + t_\eta} x_\lambda + \frac{t_\mu}{t_\lambda + t_\mu + t_\eta} x_\mu + \frac{t_\eta}{t_\lambda + t_\mu + t_\eta} x_\eta$$
$$\bar{\boldsymbol{t}} = \{\frac{T}{3}, \frac{T}{3}, \frac{T}{3}\}$$

The first line assumes that the weight given to each criterion is proportional to the time the dictator spends on it. The second that the bottom up processes are such that dictators allocate equal time to all criteria if they don't distort their attention. The key feature of these assumptions is that they treat the three criteria in the same way and assure that our result is not due to us considering one of the criteria as special.

Assumption 3 (Simplifying assumptions about the attention manipulation)**.** *The attention manipulation restricts the attention to the information relevant for criterion $k \in w, \eta, g$ to $t_k = \hat{t}$. In addition:*

- *The attention manipulation is always binding*
- *The attention the dictator is forced to divert from criterion k, that is $t_k - \hat{t}$, is equally split among the other two criteria.*

This assumption introduces an attention manipulation similar to the one we used in the experiment but simpler to analyse. In the experiment, we introduced a tighter constraint on one criterion and a softer one on another criterion. Here, for simplicity, we assume that the manipulation only constrains one criterion. Moreover, the first bullet point excludes cases in which the manipulation is not binding. The second bullet point assumes that the dictators are forced to equally split among the other two criteria the attention that they have to redirect. This mechanical redirection of attention simplifies our analysis considerably because spares us from analysing how the dictators re-optimize their attention under the attention restriction. Yet, the resulting vector of attention is likely close to what we would have obtained without this assumption. $C(\boldsymbol{t}, \bar{\boldsymbol{t}})$ is increasing and convex pushing the dictators to reallocate the attention more or less evenly across the two other criteria.

Assumption 4. *[Interior solution] Both with and without the manipulation, dictators keep less than the entire pie.*

This assumption reflects dictators' behavior in our experiment: the dictator keeps the entire pie in only 2.8% of the trials.

To state the next proposition we need some additional notation. Let's define x_1^{A-Mer} the amount the Advantage dictator keeps if her attention to the merit information is restricted to $\hat{t}$. Instead, let's define x_1^{A-Lib} the share she keeps if we restrict to $\hat{t}$ her attention to the outcome information. x_1^{D-Mer} and x_1^{D-Lib} indicate the behavior of the Disadvantaged dictator under the

same restrictions. The combined effect of the attention manipulations for an Advantaged dictator is then given by $|x_1^{A-Mer} - x_1^{A-Lib}|$. For a Disadvantaged it is $|x_1^{D-Mer} - x_1^{D-Lib}|$.

Proposition 3 (The differential effect of the attention manipulation on Advantaged and Disadvantaged dictators)**.** *Under the Assumptions 1, 2, 3 and 4 the combined effect of the attention manipulations is larger for the Advantaged dictators, that is* $|x_1^{A-Mer} - x_1^{A-Lib}| > |x_1^{D-Mer} - x_1^{D-Lib}|$*, if and only if* $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$.

Remember that L and H are the piece rates per unit of effort for the Disadvantaged and Advantaged member of the pair respectively. This result obtains as a combination of two effects. When the $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$ is satisfied the Disadvantaged are not putting too much more effort in the task than the Advantaged. The first consequence of this condition is that Advantaged dictators have a larger incentive to distort their attention towards the libertarian criterion than the Disadvantaged dictators do to distort their attention towards the meritocratic criterion. Hence the behavioral effect of restricting attention towards the libertarian criterion for the Advantaged dictators is larger than the behavioral restricting attention towards the meritocratic criterion for the Disadvantaged dictators. The second consequence is that, when the condition is satisfied, it is optimal for the Advantaged dictators to distort their attention *away* from the egalitarian criterion, while it is optimal for the Disadvantaged dictators to distort their attention *towards* the egalitarian criterion. As a consequence, the Advantaged dictators spend more time on the meritocratic and libertarian criteria which are the ones affected by the attention manipulation. More time on these criteria implies that the manipulation shifts the attention and, hence, the behavior of the Advantaged dictators to a larger extent.

$\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$ is the most prevalent case in our experiment. The inequality is satisfied by 85% of trials. More importantly, for each dictator the condition is satisfied by more than 50% of the decisions they face.

A.3 Proofs

Lemmas

Before proving the propositions, it is useful to prove some lemmas that will come in handy for the following derivations.

Lemma 1 (Perfect information)**.** *Dictators always spend a positive amount of time looking at information relevant to the libertarian and meritocratic criteria. Hence, know the exact value of* w_1*,* w_2*,* e_1*, and* e_2*.*

Proof. By assumption, we know that $C(\tilde{\boldsymbol{t}}) > u(W)$ where $\tilde{\boldsymbol{t}} \ : \ t_\lambda = 0 \vee t_\mu = 0$. This means that $U(x_1, \tilde{\boldsymbol{t}}) < 0 \ \forall \ x_1 \in [0, W]$. However, the dictator can get a higher payoff if she doesn't distort

her attention. In fact, $U(x_1, \bar{\boldsymbol{t}}) = 0$. Hence the optimal vector of attention $\boldsymbol{t^*}$ is such that $t_\lambda > 0$ and $t_\mu > 0$.

By definition, t_λ is the time the dictators spend on the information about w_1 and w_2, while t_μ is the time they spend on the information about e_1 and e_2. By assumption, spending any positive time is enough to acquire the information. Hence, the dictators know the exact values of w_1, w_2, e_1, and e_2.

□

Intuitively, it is optimum for the dictator to access the perfectly informative signals about effort and output because not doing so would have an extremely high cost.

Lemma 2 (The effect of attention on $r(\boldsymbol{t})$)**.** *If the fairness criteria $k \in \{\lambda, \mu, \eta\}$ are ordered such that $k_1 > k_2 > k_3$: Then: $\dfrac{\partial r(\boldsymbol{t})}{\partial t_j} - \dfrac{\partial r(\boldsymbol{t})}{\partial t_z} > 0$ if $j > z$*

Proof. Take $j > z$ and $\boldsymbol{t}, \boldsymbol{t'} \in S \ : \ \boldsymbol{t} = \boldsymbol{t'} + \boldsymbol{\xi}$. Where $\boldsymbol{\xi} \ : \ \xi_j = \varepsilon, \xi_z = -\varepsilon, \xi_{k \neq j,z} = 0$ with $\varepsilon > 0$. By assumption, we know that $\pi_k(\boldsymbol{t})$ is increasing in t_k and decreasing in t_{-k}. Hence:

$$\begin{aligned}
&\pi_j(\boldsymbol{t}) > \pi_j(\boldsymbol{t}) \\
&\pi_z(\boldsymbol{t}) < \pi_j(\boldsymbol{t}) \\
&\pi_k(\boldsymbol{t}) = \pi_k(\boldsymbol{t}) \ \wedge \ k \neq j, z
\end{aligned}$$

We know that $r(\boldsymbol{t}) = \sum_{k \in K} \pi_k k_1$, and that $j_1 > z_1$ so it must be that $r(\boldsymbol{t}) - r(\boldsymbol{t'}) > 0$. As the last inequality must hold for every $\varepsilon > 0$ including for ε infinitesimally small, it follows that:

$$\frac{\partial r(\boldsymbol{t})}{\partial t_j} - \frac{\partial r(\boldsymbol{t})}{\partial t_z} > 0 \tag{A.2}$$

□

Intuitively, if the dictator shift attention from one criterion that posit she should keep less to one that posits she should keep more, she increases the weight she gives to the latter. As such the amount she considers fair to keep increases. This intuition holds also for infinitesimal amounts.

Corollary 2.1. *For Advantaged $\dfrac{\partial r(\boldsymbol{t})}{\partial t_\lambda} - \dfrac{\partial r(\boldsymbol{t})}{\partial t_\mu} > 0$, for disadvantaged $\dfrac{\partial r(\boldsymbol{t})}{\partial t_\lambda} - \dfrac{\partial r(\boldsymbol{t})}{\partial t_\mu} < 0$*

Proof.

$$\begin{aligned}
&sign\left[x_\lambda - x_\mu\right] = \\
=&sign\left[e_1 l_1 - \frac{(e_1 l_1 + e_2 l_2) e_1}{e_1 + e_2}\right] \\
=&sign\left[e_1 e_2 (l_1 - l_2)\right]
\end{aligned}$$

Since $e_1, e_2 > 0$ by assumption, the sign is positive if $l_1 > l_2$ and negative if $l_1 < l_2$ By definition Advantaged dictators are those for whom $l_1 > l_2$ and, vice versa, Disadvantaged dictators are those for whom $l_1 > l_2$. Hence, by Lemma 2 we conclude that for Advantaged dictators $\frac{\partial r(\boldsymbol{t})}{\partial t_\lambda} - \frac{\partial r(\boldsymbol{t})}{\partial t_\mu} > 0$, while for Disadvantaged ones $\frac{\partial r(\boldsymbol{t})}{\partial t_\lambda} - \frac{\partial r(\boldsymbol{t})}{\partial t_\mu} < 0$. □

Lemma 3 (Keeping more than the fair share)**.** *In the optimum, the dictator keeps more than the share she considers fair. That is* $x_1^*(\boldsymbol{t^*}) > r(\boldsymbol{t^*})$.

Proof. The dictator chooses $\boldsymbol{t}$ and x_1 sequentially. So at the moment of choosing x_1, $\boldsymbol{t}$ is fixed and the cost of attention are sunk. As such the maximization problem for x_1 is the following.

$$x_1^*(\boldsymbol{t}) = \underset{x_1 \in [0;w]}{\arg\max} [u(x_1) - g(x_1 - r)]$$

As the maximand is defined on a closed and bounded interval, the Weierstrass theorem assures the existence of a solution. To characterize the solution for this problem let's take the first derivative of the maximand:

$$\frac{\partial u(x_1)}{\partial x_1} = \frac{\partial g(x_1 - r)}{\partial x_1} \tag{A.3}$$

The LHS indicates the marginal benefit of increasing x_1, this benefit is strictly positive because $u(x_1)$ is increasing. The RHS, instead, represent the marginal cost of increasing x_1 due to guilt. $g(x_1 - r) = 0$ if $x_1 \leq r(\boldsymbol{t})$ and positive and increasing elsewhere. Hence the marginal cost is equal to 0 if $x_1 \leq r(\boldsymbol{t})$ while it is positive if $r(\boldsymbol{t}) > x_1$.

Since $g(x_1 - r)$ is twice differentiable, it follows that it's first derivative is continuous. Hence:

$$\lim_{x_1 \to r(\boldsymbol{t})^-} \frac{\partial g(x_1 - r)}{\partial x_1} = \frac{\partial g(r - r)}{\partial x_1} = \lim_{x_1 \to r(\boldsymbol{t})^+} \frac{\partial g(x_1 - r)}{\partial x_1} = 0$$

Because $\lim_{x_1 \to r(\boldsymbol{t})^-} \frac{\partial g(x_1 - r)}{\partial x_1} = 0$. As a consequence $x_1^*(\boldsymbol{t}) > r(\boldsymbol{t})$. □

The lemma derives from the fact that keeping a bit more than what the dictator think is fair generates first order gains but only infinitesimal costs.

Corollary 3.1. *The dictator keeps strictly more than zero. That is* $x_1 > 0$

Proof. $r(\boldsymbol{t}) \geq 0$ because $r(\boldsymbol{t})$ is a linear combination of positive numbers. Hence $x_1 > r(\boldsymbol{t}) \geq 0$. □

Lemma 4 (The relationship between r and x_1)**.** *If* $x_1^* < W$*, then* $\frac{\partial x_1^*}{\partial r} > 0$*. That is, unless the dictator is already keeping everything, the share she keeps is increasing in the share she considers fair to keep.*

Proof. If we assume that $x_1^* < W$, then by Corollary 3.1, $0 < x_1 < W$. Hence, Equation A.3 above gives us the necessary and sufficient condition for x_1^*. The sufficiency of the condition follows from the concavity of $u(x_1) - g(x_1 - r)$. In fact, $u(x_1)$ is concave and $g(x_1 - r)$ is convex by assumption.

We can now differentiate Equation A.3 w.r.t. r and obtain:

$$\frac{\partial^2 u(x_1)}{\partial^2 x_1}\frac{\partial x_1}{\partial r} = \frac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r)}\left[\frac{\partial x_1}{\partial r} - 1\right] \tag{A.4}$$

$$\frac{\partial x_1}{\partial r} = \frac{-\dfrac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r)}}{\dfrac{\partial^2 u(x_1)}{\partial^2 x_1} - \dfrac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r)}} \tag{A.5}$$

$$\frac{\partial x_1}{\partial r} > 0 \tag{A.6}$$

Where the last step comes from the fact that $u(x_1)$ is concave and, by Lemma 3, $g(x_1 - r)$ is strictly convex if $x_1 > r$. □

Lemma 5. *There is an optimum* $\boldsymbol{t}^*$ *and* $\boldsymbol{t}^* \neq \bar{\boldsymbol{t}}$.

Proof. The maximization problem for $\boldsymbol{t}$ is given by:

$$\boldsymbol{t}^* = \arg\max_{\boldsymbol{t} \in S} U(x_1^*(\boldsymbol{t}), \boldsymbol{t}) = \arg\max_{\boldsymbol{t} \in S} u(x_1^*(\boldsymbol{t})) - g(x_1^*(\boldsymbol{t}) - \hat{r}_1(\boldsymbol{t})) - C(\boldsymbol{t}, \bar{\boldsymbol{t}})$$
$$\text{Subject to: } t_\lambda + t_\mu + t_\eta = T$$

As S is a closed and bounded subset of the domain of $U(x_1^*(\boldsymbol{t}), \boldsymbol{t})$, we know that there must exist an optimum $\boldsymbol{t}^* \in S$ by the Weierstrass theorem.

The FOCs of the problem are given by

$$\begin{cases} \dfrac{\partial u(x_1^*(\boldsymbol{t}))}{\partial x_1}\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\mu} - \dfrac{\partial g(x_1^*(\boldsymbol{t}) - r(\boldsymbol{t}))}{\partial (x_1 - r)}\left[\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\mu} - \dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\mu}\right] - \dfrac{\partial C(\boldsymbol{t})}{\partial t_\mu} = \nu \\ \dfrac{\partial u(x_1^*(\boldsymbol{t}))}{\partial x_1}\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\lambda} - \dfrac{\partial g(x_1^*(\boldsymbol{t}) - r(\boldsymbol{t}))}{\partial (x_1 - r)}\left[\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\lambda} - \dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\lambda}\right] - \dfrac{\partial C(\boldsymbol{t})}{\partial t_\lambda} = \nu \\ \dfrac{\partial u(x_1^*(\boldsymbol{t}))}{\partial x_1}\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\eta} - \dfrac{\partial g(x_1^*(\boldsymbol{t}) - r(\boldsymbol{t}))}{\partial (x_1 - r)}\left[\dfrac{\partial x_1(\boldsymbol{t})}{\partial t_\eta} - \dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\eta}\right] - \dfrac{\partial C(\boldsymbol{t})}{\partial t_\eta} = \nu \end{cases}$$

Where ν is the Lagrangian multiplier. If $\boldsymbol{t}^* = \bar{\boldsymbol{t}}$, the FOC above must hold in $\boldsymbol{t} = \bar{\boldsymbol{t}}$. Applying

the envelop theorem and substituiting ν, we can rewrite one of the FOCs in $\boldsymbol{t} = \bar{\boldsymbol{t}}$ as:

$$\frac{\partial g(x_1^*(\bar{\boldsymbol{t}})r(\bar{\boldsymbol{t}})}{\partial(x_1 - r(\boldsymbol{t}))}\left[\frac{\partial r(\bar{\boldsymbol{t}})}{\partial t_\mu} - \frac{\partial r(\bar{\boldsymbol{t}})}{\partial t_\lambda}\right] - \frac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\mu} + \frac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\lambda} =$$
$$\frac{\partial g(x_1^*(\bar{\boldsymbol{t}}) - r(\bar{\boldsymbol{t}})}{\partial(x_1 - r(\boldsymbol{t}))}\left[\frac{\partial r(\bar{\boldsymbol{t}})}{\partial t_\mu} - \frac{\partial r(\bar{\boldsymbol{t}})}{\partial t_\lambda}\right] \neq 0$$

The first step comes from the fact that $\dfrac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\mu} = \dfrac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\lambda} = 0$. In fact, by assumption we know that $C(\boldsymbol{t}, \bar{\boldsymbol{t}})$ is increasing in $\|\boldsymbol{t} - \bar{\boldsymbol{t}}\|$ and that $C(\boldsymbol{t}, \bar{\boldsymbol{t}})$ is differentiable. Hence, $C(\boldsymbol{t}, \bar{\boldsymbol{t}})$ has a minimum in $\bar{\boldsymbol{t}}$ and in that point the $\dfrac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\mu} = \dfrac{\partial C(\bar{\boldsymbol{t}})}{\partial t_\lambda} = 0$. The last line is different from zero because $\dfrac{\partial g(x_1^*(\boldsymbol{t}) - r(\boldsymbol{t}))}{\partial(x_1 - r)} > 0$ by Lemma 3, and $\dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\mu} - \dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\lambda} \neq 0$ by Corollary 2.1. Hence we conclude that $\boldsymbol{t}^* \neq \bar{\boldsymbol{t}}$. □

We are now ready to prove Proposition 1

Proof of Proposition 1

Proof. Define as $\boldsymbol{t^{*A}}$ and $\boldsymbol{t^{*D}}$ the optimal vector of attention for the Advantaged and the Disadvantaged dictators respectively. We want to prove that $t_\mu^{*A} - t_\lambda^{*A} < t_\mu^{*D} - t_\lambda^{*D}$. To do so, we will prove that $t_\mu^{*A} - t_\lambda^{*A} < \bar{t}_\mu - \bar{t}_\lambda < t_\mu^{*D} - t_\lambda^{*D}$. Where $\bar{t}_\mu$ and $\bar{t}_\lambda$ are the same for Advantaged and Disadvantaged dictators because, by assumption, the bottom up vector of attention $\bar{\boldsymbol{t}}$ does not depend on whether the dictator is Advantaged or Disadvantaged.

Let's first prove that $t_\mu^{*A} - t_\lambda^{*A} < \bar{t}_\mu - \bar{t}_\lambda$. A sufficient condition for this inequality to hold is:

$$(t_\mu^{*A} - \bar{t}_\mu < 0 \wedge t_\lambda^{*A} - \bar{t}_\lambda \geq 0) \vee (t_\mu^{*A} - \bar{t}_\mu \leq 0 \wedge t_\lambda^{*A} - \bar{t}_\lambda > 0) \tag{A.7}$$

To prove that expression A.7 is true we will begin showing that any $\boldsymbol{t} \ : \ t_\mu > \bar{t}_\mu, \ t_\lambda < \bar{t}_\lambda$ cannot be optimum. To do so, lets define $\boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{t''} \in S \ : \ t_\mu > \bar{t}_\mu, \ t_\lambda < \bar{t}_\lambda, \ \boldsymbol{t'} = \boldsymbol{t} + \boldsymbol{\xi}, \ \boldsymbol{t''} = \boldsymbol{t} + \boldsymbol{\xi} + \boldsymbol{\xi'}$. Where $\boldsymbol{\xi} \ : \ \xi_\lambda = 0, \ \xi_\mu = -\varepsilon, \ \xi_\eta = 0$ and $\boldsymbol{\xi'} \ : \ \xi'_\lambda = \varepsilon, \ \xi'_\mu = 0, \ \xi'_\eta = 0$ with $\varepsilon > 0$.

$$
\begin{aligned}
&sign\left[U(x_1^*(\boldsymbol{t}),\boldsymbol{t}) - U(x_1^*(\boldsymbol{t''}),\boldsymbol{t''})\right] = \\
&sign\left[U(x_1^*(\boldsymbol{t}),\boldsymbol{t}) - U(x_1^*(\boldsymbol{t'}),\boldsymbol{t'}) + U(x_1^*(\boldsymbol{t'}),\boldsymbol{t'}) - U(x_1^*(\boldsymbol{t''}),\boldsymbol{t''})\right] = \\
=&sign\left[\lim_{\varepsilon\to 0}\left(\frac{U(x_1^*(\boldsymbol{t}),\boldsymbol{t}) - U(x_1^*(\boldsymbol{t}+\boldsymbol{\xi}),\boldsymbol{t}+\boldsymbol{\xi})}{\varepsilon}\right) + \lim_{\varepsilon\to 0}\left(\frac{U(x_1^*(\boldsymbol{t}+\boldsymbol{\xi}),\boldsymbol{t}+\boldsymbol{\xi}) - U(x_1^*(\boldsymbol{t}+\boldsymbol{\xi}+\boldsymbol{\xi'}),\boldsymbol{t}+\boldsymbol{\xi}+\boldsymbol{\xi'})}{\varepsilon}\right)\right] = \\
=&sign\left[\frac{\partial U(x_1^*(t))}{\partial t_\lambda} - \frac{\partial U(x_1^*(t))}{\partial t_\mu}\right] = \\
=&sign\left[\frac{\partial g(x_1-\hat{r}_1)}{\partial(x_1-\hat{r}_1)}\frac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\lambda} - \frac{\partial C(\boldsymbol{t})}{\partial t_\lambda} - \frac{\partial g(x_1-\hat{r}_1)}{\partial(x_1-\hat{r}_1)}\frac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\mu} + \frac{\partial C(\boldsymbol{t})}{\partial t_\mu}\right] = \\
=&sign\left[\frac{\partial g(x_1-\hat{r}_1)}{\partial(x_1-\hat{r}_1)}\left(\frac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\lambda} - \frac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\mu}\right) + \frac{\partial C(\boldsymbol{t})}{\partial t_\mu} - \frac{\partial C(\boldsymbol{t})}{\partial t_\lambda}\right] = \text{Positive}
\end{aligned}
$$

Where to derive $\dfrac{\partial U(x_1^*(t))}{\partial t_\lambda}$ and $\dfrac{\partial U(x_1^*(t))}{\partial t_\mu}$ we used the envelop theorem. The sign of the expression is positive because $g(x_1 - r)$ is increasing, $\dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\lambda} - \dfrac{\partial \hat{r}_1(\boldsymbol{t})}{\partial t_\mu} > 0$ for Advantaged dictators by Corollary 2.1. Moreover, $C(.)$ is increasing in $|t_k - \bar{t}_k| \forall k \in K$ and, in $\boldsymbol{t}$, $t_\mu > \bar{t}_\mu$ while $t_\lambda < \bar{t}_\lambda$, hence $\dfrac{\partial C(\boldsymbol{t})}{\partial t_\mu} > 0$ while $\dfrac{\partial C(\boldsymbol{t})}{\partial t_\lambda} < 0$.

As such, every $\boldsymbol{t} \in S \ : \ t_\mu > \bar{t}_\mu \wedge t_\lambda < \bar{t}_\lambda$ cannot be optimal for Advantaged dictators.

Also $\boldsymbol{t} \in S \ : \ t_\mu = \bar{t}_\mu \wedge t_\lambda = \bar{t}_\lambda$ cannot be optimum. In fact, $t_\mu = \bar{t}_\mu \wedge t_\lambda = \bar{t}_\lambda$ implies $t_\eta = \bar{t}_\eta$ and, hence, $\boldsymbol{t} = \bar{\boldsymbol{t}}$. From Lemma 5, we know that $\boldsymbol{t}^* \neq \bar{\boldsymbol{t}}$.

Summing up, Expression A.7 is true because we have just excluded the complementary case $(t_\mu^{*A} - \bar{t}_\mu \geq 0 \wedge t_\lambda^{*A} - bart_\lambda \leq 0)$ and because Lemma 5 assures the existence of an optimal $\boldsymbol{t}$. As a consequence $t_\mu^{*A} - t_\lambda^{*A} < \bar{t}_\mu - \bar{t}_\lambda$.

The proof for $\bar{t}_\mu - \bar{t}_\lambda < t_\mu^{*D} - t_\lambda^{*D}$ involves showing that

$$
(\bar{t_\mu} - t_\mu^{*D} < 0 \wedge \bar{t_\lambda} - t_\lambda^{*D} \geq 0) \vee (\bar{t_\mu} - t_\mu^{*D} \leq 0 \wedge \bar{t_\lambda} - t_\lambda^{*D} > 0) \tag{A.8}
$$

As the proof follows the same steps as the proof for the Advantaged case, it is omitted for reasons of space. □

Proof of Corollary 1.1

Proof. Let's first look at the Advantaged Dictators. expression A.7 above implies $r(\boldsymbol{t}^*) > r(\bar{\boldsymbol{t}})$. Hence, by Lemma 4 $x_1^*(\boldsymbol{t}^*) > x_1^*(\bar{\boldsymbol{t}})$.

Similarly for Disadvantaged dictators, Expression A.8 above implies $r(\boldsymbol{t}^*) > r(\bar{\boldsymbol{t}})$. Hence, by Lemma 4 $x_1^*(\boldsymbol{t}^*) > x_1^*(\bar{\boldsymbol{t}})$. □

Proof of Proposition 2

Proof. Since $\Delta - \text{Attention}^{\beta} > \Delta - \text{Attention}^{\alpha}$, $t_{\mu}^{*\beta} + t_{\lambda}^{*\beta} = t_{\mu}^{*\alpha} + t_{\lambda}^{*\alpha}$, and $\sum_{k \in K} t_k^{*\beta} = \sum_{k \in K} t_k^{*\alpha}$, then:

$$(t_{\mu}^{*\beta} > t_{\mu}^{*\alpha} \wedge t_{\lambda}^{*\beta} \leq t_{\lambda}^{*\alpha} \wedge t_{\eta}^{*\beta} = t_{\eta}^{*\alpha}) \vee (t_{\mu}^{*\beta} \geq t_{\mu}^{*\alpha} \wedge t_{\lambda}^{*\beta} < t_{\lambda}^{*\alpha} \wedge t_{\eta}^{*\beta} =_{\eta}^{*\alpha})$$

The expression above implies $r(\boldsymbol{t}^{\beta}) < r(\boldsymbol{t}^{\alpha})$ for Advantaged dictators and $r(\boldsymbol{t}^{*\beta}) > r(\boldsymbol{t}^{*\alpha})$ for Disadvantaged ones by Corollary 2.1. Hence, Advantaged dictators keep more money for themselves and Disadvantaged ones keep less for themselves by Lemma 4. As a result the Advantaged member of the pair receives more money when $\boldsymbol{t} \in \hat{T}^{\alpha}$ than if $\boldsymbol{t} \in \hat{T}^{\beta}$. □

Proof of Proposition 3

Proof. We want to prove that under the Assumptions 1, 2, 3, and 4 $|x_1^{A-Mer} - x_1^{A-Lib}| > |x_1^{D-Mer} - x_1^{D-Lib}|$ if and only if $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. The proof has two parts. In the first we will rewrite $|x_1^{A-Mer} - x_1^{A-Lib}| > |x_1^{D-Mer} - x_1^{D-Lib}|$ in an expression that depends on the optimal vector of attention when the attention manipulation is not present and on the fairness criteria. The second part of the prove finds an explicit solution for the optimal vector of attention and shows that $|x_1^{A-Mer} - x_1^{A-Lib}| > |x_1^{D-Mer} - x_1^{D-Lib}|$ is positive if and only if $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. The third part of the proof provides the intuition behind the result.

First part of the proof of Proposition 3 This part of the proof is dedicated to rewrite $|x_1^{A-Mer} - x_1^{A-Lib}| > |x_1^{D-Mer} - x_1^{D-Lib}|$ into an expression that depends on the fairness criteria and the dictators' vector of attention.

Intermediate step 1. The attention manipulation shifts the dictator's vector of attention, which in turn shifts r, the amount that the dictator considers fair to keep. As a first step, let's check how x_1 changes with r.

From the proof of Lemma 4 we know that if $x_1 < W$ (which is the relevant case under Assumption 3):

$$\frac{\partial^2 u(x_1)}{\partial^2 x_1} \frac{\partial x_1}{\partial r(\boldsymbol{t})} = \frac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r(\boldsymbol{t}))} \left[\frac{\partial x_1}{\partial r(\boldsymbol{t})} - 1 \right]$$

$$\frac{\partial x_1}{\partial r(\boldsymbol{t})} = \frac{-\dfrac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r(\boldsymbol{t}))}}{\dfrac{\partial^2 u(x_1)}{\partial^2 x_1} - \dfrac{\partial^2 g(x_1 - r)}{\partial^2 (x_1 - r(\boldsymbol{t}))}}.$$

From Assumption 1 we have $u(x_1) = x_1$. Hence,

$$\frac{\partial x_1}{\partial r} = 1.$$

As such, the effect of the manipulation on r are translated one to one x_1. That is:

$$|x_1^{A-Mer} - x_1^{A-Lib}| = |r^{A-Mer} - r^{A-Lib}|.$$

Where r^{A-Mer} and r^{A-Lib} are the value of r for the Advantaged dictator after restricting the time she can spend on the Merit and Outcome information, respectively. The same is, of course, true for the Disadvantaged dictators.

Intermediate step 2. Let's check how the attention manipulation affects r. As an example, we compute $r^* - r^{A-Mer}$. By Assumption 3, the manipulation reduces the time the dictator spends on the Merit info by $t_\mu^* - \hat{t}$, and it increases the time she spends contemplating information relevant for the libertarian and egalitarian criteria by $\frac{1}{2}(t_\mu^* - \hat{t})$. This means that, both with and without the manipulation, the dictator spends a total of time equal to T looking at the information. Hence:

$$r(\boldsymbol{t^*}) - r^{A-Mer} = \frac{t_\lambda^{A*}}{T}x_\lambda^A + \frac{t_\mu^{A*}}{T}x_\mu^A + \frac{t_\eta^{A*}}{T}x_\eta - \frac{t_\lambda^{A*} + \frac{1}{2}(t_\mu^{A*} - \hat{t})}{T}x_\lambda^A - \frac{\hat{t}}{T}x_\mu^A - \frac{t_\eta^{A*} + \frac{1}{2}(t_\mu^{A*} - \hat{t})}{T}x_\eta =$$
$$= \frac{1}{T}\left(\hat{t} - t_\mu^{A*}\right)\left(+\frac{1}{2}x_\lambda^A - x_\mu^A + \frac{1}{2}x_\eta\right).$$

The term in the first parenthesis indicates the decrease in attention to the meritocratic criterion due to the manipulation. Instead, the term in the second parenthesis indicates the change in r per every unit of attention that the manipulation moves away from the meritocratic criterion.

With similar steps we obtain:

$$r(\boldsymbol{t^*}) - r^{A-Lib} = \frac{1}{T}\left(\hat{t} - t_\lambda^{A*}\right)\left(-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta\right)$$
$$r(\boldsymbol{t^*}) - r^{D-Mer} = \frac{1}{T}\left(\hat{t} - t_\mu^{D*}\right)\left(\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)$$
$$r(\boldsymbol{t^*}) - r^{D-Lib} = \frac{1}{T}\left(\hat{t} - t_w^{D*}\right)\left(-x_\lambda^D + \frac{1}{2}x_\mu^D + \frac{1}{2}x_\eta\right).$$

Intermediate step 3. We can now study the sign of $r^{A-Mer} - r^{A-Lib}$ and $r^{D-Mer} - r^{D-Lib}$ so that we can rewrite the inequality we want to prove without the absolute value. The total effect of the manipulations on the advantaged is given by:

$$r^{A-Mer} - r^{A-Lib} = [r^{A-Mer} - r(\boldsymbol{t^*})] - [r^{A-Lib} - r(\boldsymbol{t^*})] =$$
$$= \frac{1}{T}\left(t_\lambda^{A*} - \hat{t}\right)\left(x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta\right) - \frac{1}{T}\left(t_\mu^{A*} - \hat{t}\right)\left(-\frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta\right) > 0.$$

The last inequality comes from the fact that by Assumption 3 $\bar{t_\lambda} = \bar{t_\mu} = T/3$ and that by the proof of Proposition 1 $t_\mu^{A*} - t_\lambda^{A*} < \bar{t_\lambda} - \bar{t_\mu}$. Hence $\left(t_\lambda^{A*} - \hat{t}\right) > \left(t_\mu^{A*} - \hat{t}\right)$. Moreover, $x_\lambda^A > x_\mu^A$ and so $\left(x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta\right) > \left(-\frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta\right)$.

Using the same steps as above, one can easily prove that:

$$r^{D-Mer} - r^{D-Lib} = [r^{D-Mer} - r(\boldsymbol{t^*})] - [r^{D-Lib} - r(\boldsymbol{t^*})] =$$
$$= \frac{1}{T}\left(t_w^{D*} - \hat{t}\right)\left(x_\lambda^D - \frac{1}{2}x_\mu^D - \frac{1}{2}x_\eta\right) - \frac{1}{T}\left(t_\mu^{D*} - \hat{t}\right)\left(-\frac{1}{2}x_\lambda^D + x_\mu^D - \frac{1}{2}x_\eta\right) < 0$$

Putting together the steps so far. Using the results of the intermediate steps above we can rewrite:

$$\begin{aligned}
&|x_1^{A-Mer} - x_1^{A-Lib}| - |x_1^{D-Mer} - x_1^{D-Lib}| = \\
=&\frac{1}{T}\left(t_\lambda^{A*} - \hat{t}\right)\left(x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta\right) - \frac{1}{T}\left(t_\mu^{A*} - \hat{t}\right)\left(-\frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta\right) + \\
&+ \frac{1}{T}\left(t_w^{D*} - \hat{t}\right)\left(x_\lambda^D - \frac{1}{2}x_\mu^D - \frac{1}{2}x_\eta\right) - \frac{1}{T}\left(t_\mu^{D*} - \hat{t}\right)\left(-\frac{1}{2}x_\lambda^D + x_\mu^D - \frac{1}{2}x_\eta\right) \\
=&\frac{1}{T}(t_\lambda^{A*} - t_\lambda^{D*} + t_\lambda^{D*})\left(x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta\right) + \frac{1}{T}t_\lambda^{D*}\left(x_\lambda^D - \frac{1}{2}x_\mu^D - \frac{1}{2}x_\eta\right) + \\
&- \frac{1}{T}\hat{t}\left(x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta + x_\lambda^D - \frac{1}{2}x_\mu^D - \frac{1}{2}x_\eta\right) + \\
&- \frac{1}{T}\left(t_\mu^{D*} - t_\mu^{A*} + t_\mu^{A*}\right)\left(-\frac{1}{2}x_\lambda^D + x_\mu^D - \frac{1}{2}x_\eta\right) - \frac{1}{T}t_\mu^{A*}\left(-\frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta\right) + \\
&+ \hat{t}\left(-\frac{1}{2}x_\lambda^D + x_\mu^D - \frac{1}{2}x_\eta - \frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta\right)
\end{aligned}$$

$x_\lambda^A + x_\lambda^D = W$ and $x_\mu^A + x_\mu^D = W$. As such $-\frac{1}{2}x_\lambda^D + x_\mu^D - \frac{1}{2}x_\eta - \frac{1}{2}x_\lambda^A + x_\mu^A - \frac{1}{2}x_\eta = 0$ and $x_\lambda^A - \frac{1}{2}x_\mu^A - \frac{1}{2}x_\eta + x_\lambda^D - \frac{1}{2}x_\mu^D - \frac{1}{2}x_\eta = 0$. Hence we can rewrite the last expression as:

$$\frac{1}{T}\left[(t_\lambda^{D*} - t_\lambda^{A*})\left(-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta\right) - \left(t_\mu^{A*} - t_\mu^{D*}\right)\left(+\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)\right] \tag{A.9}$$

The first product in A.9 is the reduction in the amount the Advantaged dictators keep when they are forced to look at the libertarian criterion for the amount of time that is optimal for the Disadvantaged dictators. Vice versa, the second product indicates the reduction in the amount the Disadvantaged dictators when they are forced to look at the meritocratic criterion for the amount of time that is optimal for the Advantaged dictators. Hence, to prove that the behavioral effect of our attention manipulation is stronger for the Advantaged dictators, we need to show that the first effect is stronger than the second one: A.9 must be larger than zero.

Second part of the proof of Proposition 3 To check when Expression A.9 is positive we need first to solve the dictator maximization problem to find an explicit solution for $\boldsymbol{t}^*$.

Finding the optimal vector of attention $\boldsymbol{t}^*$. Under Assumption 1 and 2 the dictator's utility function is:

$$\begin{aligned} U(x,\boldsymbol{t}) =& x_1 - \frac{\beta}{2}(x_1 - \left(\frac{t_\lambda}{t_\lambda+t_\mu+t_\eta}x_\lambda + \frac{t_\mu}{t_\lambda+t_\mu+t_\eta}x_\mu + \frac{t_\eta}{t_\lambda+t_\mu+t_\eta}x_\eta\right)^2 + \\ & - \frac{\gamma}{2}\left[(t_\lambda - \bar{t_\lambda})^2 + (t_\mu - \bar{t_\mu})^2 + (t_\eta - \bar{t_\eta})^2\right] \end{aligned}$$

As before, we solve the problem sequentially. First the dictator finds the optimal amount to keep x_1^* as a function of $\boldsymbol{t}$. Then she chooses the optimal $\boldsymbol{t}^*$. So the first step is finding:

$$x_1^* = \underset{x_1\in[0;X]}{\arg\max}\, x_1 - \frac{\beta}{2}\left(x_1 - \left(\frac{t_\lambda}{t_\lambda+t_\mu+t_\eta}x_\lambda + \frac{t_\mu}{t_\lambda+t_\mu+t_\eta}x_\mu + \frac{t_\eta}{t_\lambda+t_\mu+t_\eta}x_\eta\right)\right)^2$$

The first order conditions is given by

$$1 = \frac{\beta(x_\eta t_\eta + x_\lambda t_\lambda + x_\mu t_\mu - t_\eta x - t_\lambda x - t_\mu x)}{t_\eta + t_\lambda + t_\mu}$$

From which, noting that the maximand is concave and that the problem has an interior solution according to Assumption 3:

$$x_1^* = \frac{t_\eta + t_\lambda + t_\mu + \beta x_\eta t_\eta + \beta x_\lambda t_\lambda + \beta x_\mu t_\mu}{\beta(t_\eta + t_\lambda + t_\mu)}$$

We can now feed the expression for x_1^* in the utility function and find $\boldsymbol{t}^*$.

$$\begin{aligned} \boldsymbol{t}^* = & \underset{\boldsymbol{t}^*\in[0;T]x[0;T]x[0;T]}{\arg\max}\; x_1^* - \frac{\beta}{2}(x_1^* - \left(\frac{t_\lambda}{t_\lambda+t_\mu+t_\eta}x_\lambda + \frac{t_\mu}{t_\lambda+t_\mu+t_\eta}x_\mu + \frac{t_\eta}{t_\lambda+t_\mu+t_\eta}x_\eta\right)^2 + \\ & - \frac{\gamma}{2}\left[(t_\lambda - \bar{t_\lambda})^2 + (t_\mu - \bar{t_\mu})^2 + (t_\eta - \bar{t_\eta})^2\right] \\ & s.t \\ & t_\lambda + t_\mu + t_\eta = T \end{aligned}$$

Which gives:

$$t_\lambda^* = \frac{T}{3} + \frac{2x_\lambda - x_\mu - x_\eta}{3T\gamma}, \tag{A.10}$$

$$t_\mu^* = \frac{T}{3} + \frac{-x_\lambda + 2x_\mu - x_\eta}{3T\gamma}, \tag{A.11}$$

$$t_\eta^* = \frac{T}{3} + \frac{-x_\lambda - x_\mu + 2x_\eta}{3T\gamma}. \tag{A.12}$$

The optimal time the dictator spends on a criterion is equal to the value of the bottom up vector of attention plus the attention distortion due to the top-down attention processes. The distortion for one criterion is increasing in the amount that criterion says it is fair to keep, but it is decreasing in the amount the other criteria say it is fair. Intuitively, the gains from distorting attention increase in the distance between criteria. Moreover, the weight of a criterion is equal to 2 in the solution for the optimal time spent on that criterion and -1 in the solution for the other two criteria. The sum of the weight is zero because the time added to one criterion needs to be taken from the other two; the two negative weights are equal to each other due to the symmetry of the cost of distorting attention. Finally, the size of the attention distortion decreases in the γ and in T. γ is the parameter that indicates the relative importance of the cost of manipulating attention, hence an higher γ indicates that the attention manipulation is more expensive. T enters in the solution, because the benefit of shifting one unit of attention is decreasing in the total time the dictator can look at the information.

Substituting the solution for $\boldsymbol{t^*}$ *in Expression A.9*. We can now go back to check under which conditions Expression A.9 is positive. To do so it is useful to define e^A and e^D as the effort of the Advantaged and Disadvantaged people, respectively. After substituting A.10, A.11, A.12 in Expression A.9, replacing x^A_λ, x^D_λ, x^A_μ and, x^D_μ with their definitions, and some algebra, we can rewrite Expression A.9 as:

$$\frac{e^A e^D \left[H(e^A)^2 - L(e^D)^2\right](H-L)}{T\gamma(e^A - e^D)^2} \tag{A.13}$$

Which is positive if and only if $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. This inequality tells us that Expression A.9 is positive as long as the Disadvantaged person does not put too much more effort than the advantaged one (remember $L < H$). Alternatively, we can interpret $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$ as telling us that the Advantaged are so much more productive per unit of effort that the Disadvantaged are far from closing the gap in production even when they put more effort than the Advantaged.

The last results proves Proposition 3. Yet, to get to the intuition behind the result we need more work.

Building the intuition behind Proposition 3 We will now try to understand intuitively why Proposition 3 is true. To do so, let's start considering the elements of Expression A.9 one by one.

The first product in Expression A.9 is

$$(t^{D*}_\lambda - t^{A*}_\lambda)\left(-x^A_\lambda + \frac{1}{2}x^A_\mu + \frac{1}{2}x_\eta\right)$$

Using A.10, A.11, A.12, the definitions of the fairness criteria, and some algebra, we derive that both terms in the products are positive if

$$H(e^A)^2 - L(e^D)^2 + 3(H-L)e^A e^D < 0. \tag{A.14}$$

As such the product is never negative. To see why the two terms switch sign together consider that

$$sign[-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta] = \tag{A.15}$$

$$= - sign[2x_\lambda^A - x_\mu^A - x_\eta] = \tag{A.16}$$

$$= sign[-2(X - x_\lambda^D) + (X - x_\mu^D) + (X - x_\eta)] = \tag{A.17}$$

$$= sign[2x_\lambda^D - x_\mu^D - x_\eta]. \tag{A.18}$$

In the second line, the expression in brackets is the numerator of the second term of Equation A.10 for the case in which the Dictator is *Advantaged*. When this expression is positive, the Advantaged dictators distort their attention towards the libertarian criterion. $\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta < 0$ implies that $2x_\lambda^A - x_\mu^A - x_\eta > 0$ and that these dictators distort their attention towards the libertarian criterion. Instead, in the last line the term in brackets is the numerator of the second term of Equation A.10 for the case in which the Dictator is *Disadvantaged*. Following the same logic as above, we see that $\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta < 0$ implies $2x_\lambda^D - x_\mu^D - x_\eta < 0$ and that the Disadvantaged dictators distort their attention away from the libertarian criterion. As such, the Advantaged and the Disadvantaged dictators always distort their attention towards the libertarian criterion in opposite directions. Moreover, since the bottom-up level of attention to the libertarian criterion does not depend on the dictator status, the sign of $\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta$ also determines the sign of $t_\mu^{A*} - t_\mu^{D*}$.

We can now look at the second product in A.9:

$$\left(t_\mu^{A*} - t_\mu^{D*}\right)\left(+\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right).$$

We the usual substitutions and some algebra, we derive that both terms in this product are positive if and only if

$$H(e^A)^2 - L(e^D)^2 - 3(H-L)e^A e^D > 0. \tag{A.19}$$

Hence, this product is never negative. The intuition behind this result is similar as the one for Inequality A.14. The sign of $\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta$ determines in which direction the dictators distort their attention towards the meritocratic criterion. The direction of the distortion is always opposite for the two types of dictators and hence the sign of $\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta$ determines also the sign of $t_\mu^{A*} - t_\mu^{D*}$.

We established that both products in Expression A.9 are never negative, hence a sufficient

condition for A.9 to be positive is that

$$\left(-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta\right) < \left(+\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right) \vee \left(-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta\right) < 0 \quad \text{(A.20)}$$

and that

$$t_\lambda^{D*} + t_\mu^{D*} < t_\mu^{A*} + t_\lambda^{A*} \vee t_\lambda^{D*} - t_\lambda^{A*} < 0. \quad \text{(A.21)}$$

The second part of Condition A.20 requires that the amount that the Advantage dictators keep goes down when they are forced to reduce attention to the libertarian criterion by one unit of time. Instead, the first part of the condition requires that this drop in the amount the Advantaged dictators keep is larger than the drop in the amount the Disadvantaged dictators keep when they have to divert attention away from the meritocratic criterion. In other words, restricting by one unit the Advantaged dictators attention to the libertarian criterion should have a larger behavioral effect than restricting by one unit the Disadvantaged dictators attention to the meritocratic criterion.

Instead, the first part of Condition A.21 requires that the Advantaged dictators spend more time looking at information than Disadvantaged ones. This condition is intuitive: for the behavioral effect to be larger for the Advantaged dictators, the attention manipulation should constraint their attention to a larger extent. Finally, the second part of the condition requires that the Advantaged dictators spend more time on the libertarian than on the meritocratic criterion.

Checking Condition A.20. Let's start considering the first part of A.20. With the usual substitution and some algebra, we can rewrite

$$\left(-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta\right) < \left(+\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)$$
$$\frac{H(e^A)^2 - L(e^D)^2}{2(e^A + e^D)} > 0$$

The last condition is satisfied when $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. To see the intuition behind this result, remember that $-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta$ indicates the change in the amount the Advantaged dictator keeps if we restrict by one unit the time she spends thinking about the libertarian criterion. Moreover, let's fix e^A and see how the amounts prescribed by the fairness criteria change when we increase e^D. Let's start for the Advantaged dictators. x_λ^A remains constant, as it does not depends on e^D. x_μ^A decreases, as the Advantaged can claim a smaller fraction of the money she produced thanks to her random advantaged: $\frac{\partial x_\mu^A}{\partial e^D} = -\frac{(e^A)^2(H-L)}{(e^A+e^D)^2}$. Finally, x_η increases because the total pie is going up: $\frac{\partial x_\eta}{\partial e^D} = -\frac{L}{2}$. The speed of decrease of x_μ^A approximates to zero for high levels of e^D, while the x_η always increases at a constant speed. Hence, for high enough values of e^D, the decrease in x_η trumps the increase in x_μ^A. As a consequence the effect of distorting the Advantaged

attention away from libertarian criterion moves towards zero and could even become positive. A similar logic shows that for the Disadvantaged dictators the effect of diverting their attention away from meritocratic criterion are higher for high levels of e^D. As the effects of increasing e^D go in opposite directions for the two types of dictators, there must be a single value of $\frac{e^A}{e^D}$ above witch the condition is satisfied.

We now turn our attention to the second part of Condition A.20. It is easy to prove that $-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta < 0$ when $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. From the derivations above, we know that $-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta$ is positive when Inequality A.14 is satisfied, which is not when $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. In fact, under this condition the sum of the first two terms of Inequality A.14 is never negative, while the last term of the sum is always positive ($H > L$ by assumption).

Checking Condition A.21. Let's now move to the first part of Condition A.21. With the usual substitutions and algebra, we can rewrite:

$$t_\lambda^{D*} + t_\mu^{D*} < t_\mu^{A*} + t_\lambda^{A*}$$
$$-\frac{2}{3}\frac{H(e^A)^2 - L(e^D)^2}{3T\gamma(e^A + e^D)} < 0.$$

Hence, the inequality is satisfied if and only if $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. To see why this is the condition, we need to go back to the first part of Condition A.20 and consider that

$$sign[-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta - \left(\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)] =$$
$$sign[-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta - \left(\frac{1}{2}(X - x_\lambda^A) - (X - x_\mu^A) + \frac{1}{2}(X - x_\eta)\right)] =$$
$$sign[-x_\lambda^A - x_\mu^A + 2x_\eta].$$

From Equation A.12 we know that the term in brackets in the last line indicates the sign of the attention distortion for the Advantaged dictators for the egalitarian criterion. A negative sign indicates that these dictators divert their attention towards the egalitarian criterion. When $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$, the first line is negative, and the term in brackets in the last line must be negative as well. Hence, the Advantaged dictators spend less time than $\bar{t} = \frac{T}{3}$ thinking about the egalitarian criterion.

At the same time, we can also rewrite the same condition as:

$$sign[-x_\lambda^A + \frac{1}{2}x_\mu^A + \frac{1}{2}x_\eta - \left(\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)] =$$
$$sign[-(X - x_\lambda^D) + \frac{1}{2}(X - x_\mu^D) + \frac{1}{2}(X - x_\eta) - \left(\frac{1}{2}x_\lambda^D - x_\mu^D + \frac{1}{2}x_\eta\right)] =$$
$$sign[x_\lambda^D + x_\mu^D - 2x_\eta].$$

From Equation A.12 we know that the term in brackets in the last line indicates the sign of the attention distortion for the Disadvantaged dictators for the egalitarian criterion. A positive sign indicates that these dictators divert their attention towards the egalitarian criterion. When $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$, the first line is negative, and the term in brackets in the last line must be positive. Hence, the Disadvantaged dictators spend more time than $\bar{t} = \frac{T}{3}$ thinking about the egalitarian criterion.

We can now see the intuition behind the result that Condition A.21 is satisfied if and only if $\frac{e^A}{e^D} > \sqrt{\frac{L}{H}}$. For high value of $\frac{e^A}{e^D}$ the Advantaged Dictators distort their attention away from the egalitarian criterion, while the Disadvantaged dictators distort their attention towards this criterion. All the attention that is not directed to the egalitarian criterion must be allocated to the meritocratic and libertarian ones. Hence the Advantaged dictators spend more time on those criteria than the Disadvantaged. This relationship flips for low enough values of $\frac{e^A}{e^D}$.

Putting together the different parts of the intuition. We are now ready to explain the intuition behind Proposition 3. When the Disadvantaged put less effort or at least not too much more effort than the Advantaged two things happen. First, the behavioral effect of restricting the Advantaged dictators' attention towards the libertarian criterion by one unit of time is larger than restricting the Disadvantaged dictators' attention towards the meritocratic criterion. Second, the Advantaged dictators find more optimal to spend more time on the libertarian and meritocratic criteria than the Disadvantaged dictators. Hence the attention manipulation shifts the Advantaged dictators' attention more than the Disadvantaged dictators' attention. Both effects contributes to ensure that the combined behavioral effect of the attention manipulations is larger for the Advantaged dictators.

□

Appendix B

Memory Sophistication - Appendix

B.1 Preregistration and deviations from it

The project was registered on AsPredicted.org. The preregistration document is below.

The paper deviates from the preregistration in the analysis of question 3: “Does people’s willingness to pay (WTP) for memory-enhancing tools adjust rationally to changes in incentives?". The preregistered analysis yields statistically significant results, but it is excluded from the paper because it is not adequate to answer the research question. The main reason is that the value of a memory aid depends also on the beliefs about the probability of remembering if there is no chance of receiving the memory aid. The experiment does not elicit these beliefs since the beliefs elicitation happens after the participants indicated their WTP for the computer code - the memory aid in the experiment. At that point all the participants that indicated a positive WTP have a positive probability of receiving the computer code. The preregistered analysis is replaced with the one described in Section 3.5.3.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Memory Sophisitcation (#107628)

Created: 09/22/2022 05:16 AM (PT)

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?

On the first day of the experiment, participants study a database of color-number pairs. They then learn that they will be tested on this database during the second day of the experiment and that they can win a monetary bonus if their memory is accurate.

1. In session 1, can people anticipate how good their memory during session 2 will be?
2. After completing the test in the second session, do people have accurate beliefs about their performance?
3. Does people's willingness to pay (WTP) for memory-enhancing tools adjust rationally to changes in incentives?
4. How does the answer to the questions above change with the task difficulty?

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

Measured variables:

- number of correct answers = number of correct answers in the memory test
- WTP Low Incentives = WTP for a memory aid if the incentives for accurate memory are low. It is measured with a Multiple Price List
- WTP High Incentives = WTP for a memory aid if the incentives for accurate memory are high. It is measured with a Multiple Price List
- Beliefs ex-ante = Ex-ante subjective probability of winning the bonus during the test conditional on not receiving the memory aid. It is measured with a non-incentivized question on a 0-100 scale
- Beliefs ex-post = Ex-post subjective probability of the accuracy of one answer in the test. It is measured with a non-incentivized question on a 0-100 scale. One measurement for each question in the test

Constructed variables

- fraction correct answers = number of correct answers/number of questions
- ex-ante Sophistication = (Beliefs ex-ante)/100 – fraction correct answers
- ex-post Sophistication = (Beliefs ex-post)/100 – fraction correct answers
- Sensitivity to incentives = (WTP High Incentives - WTP Low Incentives -100+ Beliefs ex-ante)/100

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

3 conditions:

1. Easy: the participants need to remember 2 color-number pairs
2. Hard many numbers: the participants need to remember 7 color-number pairs
3. Hard interference: the participants need to remember 2 color-number pairs but memorizing these numbers is harder due to an additional task that generates interference

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

1. t-test for ex-ante Sophistication=0
2. t-test for ex-post Sophistication=0
3. t-test for Sensitivity to incentives=0
4. Running the tests described above separately for the different treatments

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

For analysis 3, I exclude participants whose WTP for at least one of the two WTP measures is top-censored.

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 will collect 1000 complete observations for session 1. I will then reinvite these 1000 participants to session 2 of the experiment. Participants who do not complete session 2 are excluded from the main analysis and are not replaced with new subjects.
The participants are randomly assigned to the different treatments. Each participant has a probability of 1/3 of being in each of the treatments.

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

As of the writing of this pre-registration, the single author of this pre-registration was the only author involved in this research project

Appendix C

Correcting Consumer Misperceptions about CO2 Emissions - Appendix

C.1 Preregistration

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Information provision about CO2 emissions and meat consumption. (#92070)

Created: 03/25/2022 03:41 AM (PT)

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?
Correcting perceptions about CO2 emissions associated with meat products will affect demand for these products.
In particular, in previous work we have used data on a) misperceptions about CO2 emissions and b) willingness to pay to avoid CO2 emissions to predict the effect of providing information about the emissions. Following these predictions, we expect that providing information about CO2 emissions will have a larger negative effect on the demand for beef than on the demand for chicken.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variable is the willingness to pay (WTP) for a package of meat products. Willingness to pay is measured by an incentive compatible multiple price list mechanism.

4) How many and which conditions will participants be assigned to?
The experiment has two parts. The first part contains our main design, which is a 2x2:
·The meat package consists of either beef products (sirloin steaks) or chicken products (chicken breasts).
·Participants either obtain a scientific estimate of the emissions associated with the package ("info" treatment) or not ("no info" treatment).
These four conditions are between-subjects.

In the second part of the experiment (again a 2x2), we will ask each subject for their WTP for the alternative meat product. In the information treatment, this implies that subjects now have knowledge about both beef and chicken products ("double info" treatment).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will regress the WTP for both meat products in Part 1 of the experiment on a treatment dummy for information provision and meat type, and we will test the interaction of meat type and information provision. Our regression analysis will control for covariates like political orientation and household income.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will not exclude observations. However, we will conduct robustness checks where we exclude people who were not able to reproduce the information we gave them in the info treatments or that did not give us their address for sending the meat products.

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.
We aim at collecting 2000 observations, 500 in each treatment cell. We consider an observation collected if a participant completed the first part of the experiment.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We conduct a questionnaire where we ask several personal characteristics. We will correlate these characteristics with WTP. We will study how people update beliefs about CO2 emissions in response to the information and will study whether prior and posterior beliefs affect purchases.
We will also conduct heterogeneity analyses by subgroups that have been shown to have a higher elasticity of meat consumption, or, per our previous survey, have shown particularly large predicted effects of information.
As robustness checks for the model specification, we will conduct Tobit regressions with censoring above. We will also look quantile regressions for 10 WTP quantiles, and focus on the interaction effects among the middle quantiles that are away from the extremes of the WTP distribution.
Finally, to understand the impact of information about substitutes, we will compare the results of the first part of the experiment (info vs. no info), with the results of the second part (double info vs. no info).

Appendix D

Other Appendices

List of co-authors and contributions

Chapter 3 of this thesis is single-authored, Chapters 1, 2 and 4 are based on co-authored work. All references of co-authored work are provided in the chapters. The contributions of the individual authors in the co-authored chapters are outlined below.

Chapter 1: “Fair Shares and Selective Attention".

Co-authors: Dianna Amasino, Joël van der Weele.

The authors developed the research idea and the experimental design jointly. Dianna and Davide programmed the experiment. Davide led the data analysis and wrote the theoretical model. All the authors contributed to writing and revising the manuscript.

Chapter 2: “Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms".

Co-authors: Dianna Amasino, Joël van der Weele.

The authors developed the research idea and the experimental design jointly. Dianna and Davide programmed the experiment. Davide led the data analysis. Joël wrote the theoretical model. All the authors contributed to writing and revising the manuscript.

Chapter 4: “Correcting Consumer Misperceptions about CO2 Emissions".

Co-authors: Taisuke Imai, Joël van der Weele, Peter Schwardmann.

The authors developed the research idea. Davide led the design of the experiment and he programmed the survey and the experiment. Taisuke led the data analysis. All the authors contributed to writing and revising the manuscript.

English summary

This thesis comprises of four chapters. Below I summarize each of them

Chapter 1: Fair Shares and Selective Attention. Attitudes towards fairness and redistribution differ along socio-economic lines, resulting in political conflict. To understand their formation, we conduct a large-scale experiment on attention to merit and luck and the effect of attention on fairness decisions. Randomly advantaged subjects pay less attention to information about true merit and retain more economic surplus, and this effect persists in subsequent impartial decisions. Attention also has a causal role: encouraging subjects to look at merit reduces the effect of an advantaged position on allocations. This suggests that attention-based policy interventions may be effective in reducing polarized views on inequality.

Chapter 2: Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms. We explore the psychological mechanisms underlying self-serving redistribution decisions in an experimental setting. This self-serving bias in redistribution has been attributed not only to self-interest, but also to constructs such as differing beliefs about the hard work or luck underlying inequality, differing fairness views, and differing perceptions of social norms. In this study, we directly measure each of these potential mechanisms and compare their mediating roles in the relationship between status and redistribution. In our experiment, participants complete real-effort tasks and then are randomly assigned a high or low pay rate per correct answer to exogenously induce (dis)advantaged status. Participants are then paired and those assigned the role of dictator decide how to divide their joint earnings. We find that advantaged dictators keep more for themselves than disadvantaged dictators and report different fairness views and beliefs about task performance, but not different perceptions of social norms. Further, only fairness views play a significant mediating role between status and allocation differences, suggesting this is the primary mechanism underlying self-serving differences in support for redistribution.

This Chapter builds on the same experimental design and the same dataset as Chapter 1.

Chapter 3: Memory Sophistication. Human memory is less than perfect, leading to mistakes in judgments and decisions. This project experimentally investigates a) whether people are sophisticated about their memory limitations, and b) whether the complexity of the memory task affects sophistication. It finds that people can be both under and overconfident about their memory. Interference makes people overconfident. A high amount of information to remember makes them underconfident. Moreover, with time confidence goes up mitigating underconfidence but exacerbating overconfidence. These findings show that memory sophistication is a complicated phenomenon, they indicate when memory limitations generate expensive mistakes, and they suggest how to minimize the cost of memory mistakes in the workplace.

Chapter 4: Correcting Consumer Misperceptions about CO2 Emissions. Policy makers put great emphasis on the role of information about carbon emissions in achieving sustainable

decisions by consumers. We conduct two studies to understand the effect of such information on consumption. First, we elicit belief distributions over the climate impact of different consumption behaviors as well as the willingness to mitigate CO_2 emissions in a representative sample of US consumers ($N = 1,022$) and combine them in a structural model to derive sharp predictions about where to best target information. In a field experiment with actual consumption decisions ($N = 2,081$), we then test for the effect of CO_2 information on the demand for beef, a product predicted to be a productive target for information. Correcting misperceptions has no effect on the demand for beef, both in absolute terms and compared to a predictably less productive target of information, i.e. the demand for poultry. Our experimental design allows us to hone in on the underlying reason for this null effect. Our results call into question the potential of even carefully-targeted information to affect individual climate action and have implications for the literature on CO_2 misperceptions and labeling.

Dutch summary

Titel: Essays over de cognitieve grondslagen van menselijk gedrag en over de gedrags economie van klimaatverandering

Hoofdstuk 1: Fair Shares and Selective Attention. De houding tegenover de rechtvaardigheid en herverdeling verschilt tussen sociaal-economische groepen, wat kan leiden tot politieke conflicten. Om deze houding beter te begrijpen voeren we een grootschalig experiment uit over aandacht voor verdiensten door hard werk en toeval, en het effect van die aandacht op rechtvaardige beslissingen. Willekeurig bevoordeelde proefpersonen besteden minder aandacht aan informatie over verdiensten en behouden meer economisch surplus voor zichzelf; dit effect blijft bestaan in latere onpartijdige beslissingen. Aandacht speelt ook een causale rol: door proefpersonen aan te moedigen naar verdienste te kijken, vermindert het effect van een bevoordeelde positie op de verdeling van het surplus. Dit suggereert effectiviteit van op aandacht gebaseerde beleidsinterventies bij het verminderen van gepolariseerde opvattingen over ongelijkheid.

Hoofdstuk 2: Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms. We onderzoeken de psychologische mechanismen die ten grondslag liggen aan "self-serving bias" in herverdelingsbeslissingen in een experimentele setting. In de literatuur wordt dit fenomeen niet alleen toegeschreven aan eigenbelang, maar ook aan psychologische verschuivingen in de rol van hard werk of toeval dat ten grondslag ligt aan ongelijkheid, verschillende opvattingen over rechtvaardigheid en verschillende percepties van sociale normen. In deze studie meten we elk van deze potentiële mechanismen rechtstreeks en vergelijken we hun bemiddelende rol in de relatie tussen status en herverdeling. In ons experiment verdienen deelnemers geld door het uitvoeren van cognitieve taken, en krijgen willekeurig een hoog of laag loon per correct antwoord toegewezen. Daarmee induceren we experimenteel een (on)bevoordeelde status aan de deelnemers. Vervolgens worden de deelnemers aan elkaar gekoppeld; degenen die de rol van "dictator" krijgen toebedeeld beslissen hoe zij hun gezamenlijke inkomsten verdelen. Bevoordeelde dictatoren behouden een groter deel van het surplus voor zichzelf dan benadeelde dictatoren, rapporteren andere persoonlijke opvattingen over rechtvaardigheid en inschattingen van relatieve prestaties, maar geen verschillende opvattingen over sociale normen van rechtvaardigheid. Alleen de persoonlijke rechtvaardigheidsopvattingen spelen een belangrijke bemiddelende rol tussen status en verdelingsbeslissingen, wat suggereert dat dit het primaire mechanisme is dat ten grondslag ligt aan "self-serving bias".

Dit hoofdstuk is gebaseerd op dezelfde experimentele opzet en dezelfde dataset als hoofdstuk

Hoofdstuk 3: Memory Sophistication. Het menselijk geheugen is niet perfect, wat leidt tot fouten in oordelen en beslissingen. In dit project wordt experimenteel onderzocht a) of mensen zich bewust zijn van hun geheugenbeperkingen, en b) of de complexiteit van de geheugentaak van invloed is op dit bewustzijn. Het blijkt dat proefpersonen zowel te weinig als te

veel vertrouwen kunnen hebben in hun geheugen: Interferentie zorgt voor overmoedigheid terwijl een groter aantal te onthouden gegevens ze te pessimistisch maakt. Bovendien neemt het zelfvertrouwen met de tijd toe, waardoor overmoedigheid toeneemt en pessimisme afneemt. Deze bevindingen tonen aan dat de inschatting van ons eigen geheugen een ingewikkeld fenomeen is, geven aan wanneer geheugenbeperkingen dure fouten veroorzaken, en suggereren hoe de kosten van geheugenfouten op de werkplek kunnen worden geminimaliseerd.

Hoofdstuk 4: Correcting Consumer Misperceptions about CO2 Emissions. Beleidsmakers leggen grote nadruk op de rol van informatie over CO2 emissies bij het nemen van duurzame beslissingen door consumenten. Wij voeren twee studies uit om het effect van dergelijke informatie op de consumptie te begrijpen. Eerst meten we in een representatieve steekproef van Amerikaanse consumenten ($N = 1.022$) de inschatting van het klimaateffect van verschillende consumptiegedragingen en de bereidheid om CO2 emissies te beperken. We combineren deze variabelen in een structureel model om voorspellingen te doen over de optimale toepassing van voorlichting. In een experiment met daadwerkelijke consumptiebeslissingen ($N = 2.081$) testen we vervolgens het effect van informatie over de CO2-uitstoot op de vraag naar rundvlees, een product waarvoor we voorspellen dat voorlichting het meest productief is. Het corrigeren van verkeerde voorstellingen heeft geen effect op de vraag naar rundvlees, zowel in absolute termen als in vergelijking met de vraag naar kippenvlees, waar informatie theoretisch een kleinere rol zou moeten spelen. Ons experiment stelt ons in staat de onderliggende reden voor dit nuleffect op te sporen. De resultaten zetten vraagtekens bij het potentieel van zorgvuldig gerichte informatie om individuele klimaatactie te beïnvloeden en hebben implicaties voor de literatuur over mispercepties en het gebruik van klimaat labels.

Link to the online Appendix

The online appendix can be found at the link below:

https://drive.google.com/file/d/1DcoAVis9D48PX3UpzGVi7Y33v$_{T}qTqGw/view$

Bibliography

Ajzen, Icek, and Martin Fishbein. 1970. "The prediction of behavior from attitudinal and normative variables." *Journal of Experimental Social Psychology*, 6(4): 466–487.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence From the Lightbulb Market." *American Economic Review*, 105(8): 2501–2538.

Almås, Ingvild, Alexander W Cappelen, Erik Ø Sørensen, and Bertil Tungodden. 2022. "Global evidence on the selfish rich inequality hypothesis." *Proceedings of the National Academy of Sciences*, 119(3).

Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden. 2020. "Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?" *Journal of Political Economy*, 128(5): 1753–1788. https://doi.org/10.1086/705551.

Amasino, Dianna, Davide Domenico Pace, and Joel van der Weele. 2023. "Self-serving Bias in Redistribution Choices: Accounting for Beliefs and Norms." https://doi.org/10.5282/ubm/epub.94539. Volume: 380.

Amasino, Dianna, Davide Pace, and Joel J van der Weele. 2021. "Fair shares and selective attention."

Ameriks, John, Andrew Caplin, John Leahy, and Tom Tyler. 2007. "Measuring Self-Control Problems." *American Economic Review*, 97(3): 966–972. https://doi.org/10.1257/aer.97.3.966.

Andreoni, James, and John Miller. 2002. "Giving according to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737–753.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2021. "Fighting Climate Change: The Role of Norms, Preferences, and Moral Values." CESifo Working Paper No. 9175.

Armel, K. Carrie, Aurelie Beaumel, and Antonio Rangel. 2008. "Biasing simple choices by manipulating relative visual attention." *Judgment and Decision making*, 3(5): 396–403.

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin. 2010. "Public Perceptions of Energy Consumption and Savings." *Proceedings of the National Academy of Sciences*, 107(37): 16054–16059.

Augenblick, Ned, and Matthew Rabin. 2019. "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *The Review of Economic Studies*, 86(3): 941–975. https://doi.org/10.1093/restud/rdy019.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. "Incentives in Experiments: A Theoretical Analysis." *Journal of Political Economy*, 126(4): 1472–1503.

Babcock, Linda, and George Loewenstein. 1997. "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic perspectives*, 11(1): 109–126.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *The American Economic Review*, 85(5): 1337–1343. http://www.jstor.org/stable/2950993 (accessed 2018-01-27).

Ba, Cuimin. 2022. "Robust Misspecified Models and Paradigm Shifts."

Barron, Kai, and Tilman Fries. 2022. "Narrative Persuasion."

Bašić, Zvonimir, and Eugenio Verrina. 2021. "Personal Norms — and Not Only Social Norms — Shape Economic Behavior." Social Science Research Network SSRN Scholarly Paper ID 3720539. https://papers.ssrn.com/abstract=3720539 (accessed 2022-01-31).

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review*, 96(5): 1652–1678. http://www.nber.org/papers/w11535.

Bernard, René, Panagiota Tzamourani, and Michael Weber. 2022. "Climate Change and Individual Behavior." Deutsche Bundesbank Discussion Paper No. 01/2022.

Bernheim, B. Douglas, and Raphael Thomadsen. 2005. "Memory and Anticipation." *The Economic Journal*, 115(503): 271–304. https://doi.org/10.1111/j.1468-0297.2005.00989.x.

Bianchi, Filippo, Claudia Dorsel, Emma Garnett, Paul Aveyard, and Susan A. Jebb. 2018. "Interventions Targeting Conscious Determinants of Human Behaviour to Reduce the Demand for Meat: A Systematic Review with Qualitative Comparative Analysis." *International Journal of Behavioral Nutrition and Physical Activity*, 15(1): 1–25.

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.

Bieleke, Maik, David Dohmen, and Peter M. Gollwitzer. 2020. "Effects of social value orientation (SVO) and decision mode on controlled information acquisition—A Mouselab perspective." *Journal of Experimental Social Psychology*, 86: 103896.

Bilén, David. 2022. "Do Carbon Labels Cause Consumers to Reduce Their Emissions? Evidence from a Large-Scale Natural Experiment." Mimeo, Gothenburg University.

Billeter, Darron, Ajay Kalra, and George Loewenstein. 2011. "Underpredicting Learning after Initial Experience with a Product." *Journal of Consumer Research*, 37(5): 723–736. https://doi.org/10.1086/655862.

Blouin, Arthur, and Sharun W. Mukand. 2019. "Erasing Ethnicity? Propaganda, Nation Building, and Identity in Rwanda." *Journal of Political Economy*, 127(3): 1008–1062. https://doi.org/10.1086/701441.

Bordalo, Pedro, John J. Conlon, Nicola Gennaioli, Spencer Y. Kwon, and Andrei Shleifer. 2023. "Memory and probability." *The Quarterly Journal of Economics*, 138(1): 265–311. Publisher: Oxford University Press.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "2Salience in Experimental Tests of the Endowment Effect." *American Economic Review*, 102(3): 47–52. https://doi.org/10.1257/aer.102.3.47.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *The Quarterly Journal of Economics*, 135(3): 1399–1442. https://doi.org/10.1093/qje/qjaa007.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2021. "Salience." National Bureau of Economic Research Working Paper 29274, https://doi.org/10.3386/w29274.

Bortolotti, Stefania, Ivan Soraperra, Matthias Sutter, and Claudia Zoller. 2017. "Too Lucky to Be True - Fairness Views Under the Shadow of Cheating." Social Science Research Network SSRN Scholarly Paper ID 3014734. https://papers.ssrn.com/abstract=3014734 (accessed 2020-09-22).

Bradley, Gifford W. 1978. "Self-serving biases in the attribution process: A reexamination of the fact or fiction question." *Journal of Personality and Social Psychology*, 36(1): 56.

Bronchetti, Erin T., Judd B. Kessler, Ellen B. Magenheim, Dmitry Taubinsky, and Eric Zwick. 2022. "Is Attention Produced Optimally? Theory and Evidence from Take-Up of Bandwidth Enhancements."

Brunner, Florentine, Verena Kurz, David Bryngelsson, and Fredrik Hedenus. 2018. "Carbon Label at a University Restaurant–Label Implementation and Evaluation." *Ecological Economics*, 146: 658–667.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029. https://doi.org/10.1257/aer.20180975.

Buser, Thomas, Gianluca Grimalda, Louis Putterman, and Joël van der Weele. 2020. "Overconfidence and gender gaps in redistributive preferences: Cross-Country experimental evidence." *Journal of Economic Behavior & Organization*, 178: 267–286. https://doi.org/10.1016/j.jebo.2020.07.005.

Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." *The Quarterly Journal of Economics*, 117(3): 871–915.

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

Camerer, Colin, and Dan Lovallo. 1999. "Overconfidence and Excess Entry: An Experimental Approach." *American Economic Review*, 89(1): 306–318. https://doi.org/10.1257/aer.89.1.306.

Camilleri, Adrian R., Richard P. Larrick, Shajuti Hossain, and Dalia Patino-Echeverri. 2019. "Consumers Underestimate the Emissions Associated with Food but Are Aided by Labels." *Nature Climate Change*, 9(1): 53–58.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø Sørensen, and Bertil Tungodden. 2007. "The pluralism of fairness ideals: An experimental approach." *American Economic Review*, 97(3): 818–827.

Cappelen, Alexander W., Erik Ø. Sørensen, and Bertil Tungodden. 2010. "Responsibility for what? Fairness and individual responsibility." *European Economic Review*, 54(3): 429–441. https://doi.org/10.1016/j.euroecorev.2009.08.005.

Cappelen, Alexander W., James Konow, Erik Ø Sørensen, and Bertil Tungodden. 2013. "Just luck: An experimental study of risk-taking and fairness." *American Economic Review*, 103(4): 1398–1413.

Cappelen, Alexander W., Karl O. Moene, Siv-Elisabeth Skjelbred, and Bertil Tungodden. 2017. "The Merit Primacy Effect." Social Science Research Network SSRN Scholarly Paper ID 2963504. https://papers.ssrn.com/abstract=2963504 (accessed 2020-09-22).

Carlsson, Fredrik, Mitesh Kataria, and Elina Lampi. 2022. "How Much Does It Take? Willingness to Switch to Meat Substitutes." *Ecological Economics*, 193: 107329.

Cassar, Lea, and Arnd H. Klein. 2019. "A Matter of Perspective: How Failure Shapes Distributive Preferences." *Management Science*, 65(11): 5050–5064. https://doi.org/10.1287/mnsc.2018.3185.

Charness, Gary, Uri Gneezy, and Vlastimil Rasocha. 2021. "Experimental methods: Eliciting beliefs." *Journal of Economic Behavior & Organization*, 189: 234–256.

Chater, Nick, and George Loewenstein. 2022. "The i-Frame and the s-Frame: How Focusing on Individual-Level Solutions Has Led Behavioral Public Policy Astray." *Behavioral and Brain Sciences*.

Chen, Si, and Carl Heese. 2021. "Motivated Information Acquisition."

Cherry, Todd L, Peter Frykblom, and Jason F Shogren. 2002. "Hardnose the dictator." *American Economic Review*, 92(4): 1218–1221.

Chew, Soo Hong, Wei Huang, and Xiaojian Zhao. 2020. "Motivated False Memory." *Journal of Political Economy*, 128(10): 3913–3939. https://doi.org/10.1086/709971.

Cialdini, Robert B, Carl A Kallgren, and Raymond R Reno. 1991. "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior." In *Advances in experimental social psychology*. Vol. 24, 201–234. Elsevier.

Clogg, Clifford C., Eva Petkova, and Adamantios Haritou. 1995. "Statistical Methods for Comparing Regression Coefficients Between Models." *American Journal of Sociology*, 100(5): 1261–1293. https://doi.org/10.1086/230638.

Cohn, Alain, Lasse J. Jessen, Marko Klasnja, and Paul Smeets. 2019. "Why Do the Rich Oppose Redistribution? An Experiment with America's Top 5%." Social Science Research Network SSRN Scholarly Paper ID 3395213. https://papers.ssrn.com/abstract=3395213 (accessed 2021-05-04).

Conlon, John J., Malavika Mani, Gautam Rao, Matthew W. Ridley, and Frank Schilbach. 2022. "Not Learning from Others." National Bureau of Economic Research.

Crawford, Vincent P., and Nagore Iriberri. 2007. "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games." *American Economic Review*, 97(5): 1731–1750. https://doi.org/10.1257/aer.97.5.1731.

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.

d'Adda, Giovanna, Yu Gao, and Massimo Tavoni. 2022. "A Randomized Trial of Energy Cost Information Provision Alongside Energy-Efficiency Classes for Refrigerator Purchases." *Nature Energy*, 7(4): 360–368.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33(1): 67–80.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review*, 112: 2851–2883.

Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. 2022. "Fighting Climate Change: International Attitudes Toward Climate Policies." NBER Working Paper No. 30265.

Deffains, Bruno, Romain Espinosa, and Christian Thöni. 2016. "Political self-serving bias and redistribution." *Journal of Public Economics*, 134: 67–74. https://doi.org/10.1016/j.jpubeco.2016.01.002.

DellaVigna, Stefano, and Ulrike Malmendier. 2004. "Contract Design and Self-Control: Theory and Evidence." *The Quarterly Journal of Economics*, 119(2): 353–402. https://doi.org/10.1162/0033553041382111.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719. https://doi.org/10.1257/aer.96.3.694.

Dengler-Roscher, Kathrin, Natalia Montinari, Marian Panganiban, Matteo Ploner, and Benedikt Werner. 2018. "On the malleability of fairness ideals: Spillover effects in partial and impartial allocation tasks." *Journal of Economic Psychology*, 65: 60–74.

DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. 2012. "How Does the Brain Solve Visual Object Recognition?" *Neuron*, 73(3): 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.

Diederich, Johannes, and Timo Goeschl. 2014. "Willingness to Pay for Voluntary Climate Action and Its Determinants: Field-Experimental Evidence." *Environmental and Resource Economics*, 57(3): 405–429.

Di Tella, Rafael, Sebastian Galiant, and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *The Quarterly Journal of Economics*, 122(1): 209–241. https://doi.org/10.1162/qjec.122.1.209.

Dong, Diansheng, Christopher G. Davis, and Hayden Stewart. 2015. "The Quantity and Variety of Households' Meat Purchases: A Censored Demand System Approach." *Agricultural Economics*, 46(1): 99–112.

Dorin, Camille, Marine Hainguerlot, Hélène Huber-Yahi, Jean-Christophe Vergnaud, and Vincent de Gardelle. 2021. "How economic success shapes redistribution: The role of self-serving beliefs, in-group bias and justice principles." *Judgment and Decision Making*, 16(4): 932.

Dunlosky, John, and Sarah Uma K. Tauber. 2016. *The Oxford handbook of metamemory.* Oxford University Press.

Durante, Ruben, Louis Putterman, and Joël Van der Weele. 2014. "Preferences for redistribution and perception of fairness: An experimental study." *Journal of the European Economic Association*, 12(4): 1059–1086.

Eckel, Catherine C, Enrique Fatas, and Malcolm Kass. 2022. "Sacrifice: An experiment on the political economy of extreme intergroup punishment." *Journal of Economic Psychology*, 90: 102486.

Ellis, Andrew, and David J Freeman. 2020. "Revealing Choice Bracketing." arXiv:2006.14869.

Elofsson, Katarina, Niklas Bengtsson, Elina Matsdotter, and Johan Arntyr. 2016. "The Impact of Climate Information on Milk Demand: Evidence from a Field Experiment." *Food Policy*, 58: 14–23.

Engelmann, Jan, Alejandro Hirmas, and Joel J van der Weele. 2021. "Top Down or Bottom Up? Disentangling the Channels of Attention in Risky Choice."

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation neglect in belief formation." *The Review of Economic Studies*, 86(1): 313–332. Publisher: Oxford University Press.

Enke, Benjamin, and Thomas Graeber. 2019. "Cognitive uncertainty." National Bureau of Economic Research.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. 2022. "Associative Memory and Beliefs formation." `https://drive.google.com/file/d/1MQFWekxbBw9VBInmrygBJSn-T_pi_YQE/view?usp=sharing&usp=embed_facebook` (accessed 2023-04-04).

Enke, Benjamin, Ricardo Rodriguez-Padilla, and Florian Zimmermann. 2022. "Moral universalism: Measurement and economic relevance." *Management Science*, 68(5): 3590–3603. Publisher: INFORMS.

Espinosa, Romain, Bruno Deffains, and Christian Thöni. 2020. "Debiasing preferences over redistribution: an experiment." *Social Choice and Welfare*, 55(4): 823–843. `https://doi.org/10.1007/s00355-020-01265-z`.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel. 2020. "Mental models and learning: The case of base-rate neglect." Tech. rep.

European Commission. 2020. "New Consumer Agenda: Strengthening Consumer Resilience for Sustainable Recovery." COM(2020) 696 final. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0696&qid=1605887353618`.

Fairbrother, Malcolm. 2022. "Public Opinion About Climate Policies: A Review and Call for More Studies of What People Want." *PLOS Climate*, 1(5): e0000030.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global evidence on economic preferences." *The Quarterly Journal of Economics*, 133(4): 1645–1692. Publisher: Oxford University Press.

Fallucchi, Francesco, and Marc Kaufmann. 2021. "Narrow Bracketing in Work Choices." arXiv:2101.04529.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and punishment in public goods experiments." *American Economic Review*, 90(4): 980–994.

Fiedler, Susann, and Adrian Hillenbrand. 2020. "Gain-loss framing in interdependent choice." *Games and Economic Behavior*, 121: 232–251.

Fiedler, Susann, Andreas Glöckner, Andreas Nicklisch, and Stephan Dickert. 2013. "Social Value Orientation and information search in social dilemmas: An eye-tracking analysis." *Organizational Behavior and Human Decision Processes*, 120(2): 272–284. https://doi.org/10.1016/j.obhdp.2012.07.002.

Fischbacher, Urs, David Grammling, and Jan Hausfeld. 2021. "Redistribution beyond equality and status quo-heterogeneous societies in the lab." Thurgauer Wirtschaftsinstitut, UniversitÃ¤ t Konstanz.

Fischbacher, Urs, Jan Hausfeld, and Baiba Renerte. 2020. "Strategic incentives undermine gaze as a signal of prosocial motives." Thurgauer Wirtschaftsinstitut, UniversitÃ¤ t Konstanz.

Fisher, Geoffrey. 2021. "Intertemporal choices are causally influenced by fluctuations in visual attention." *Management Science*.

Fisman, Raymond, Shachar Kariv, and Daniel Markovits. 2007. "Individual Preferences for Giving." *American Economic Review*, 97(5): 1858–1876.

Fosgaard, Toke Reinholt, Alice Pizzo, and Sally Sadoff. 2021. "Do People Respond to the Climate Impact of Their Behavior? The Effect of Carbon Footprint Information on Grocery Purchases." IFRO Working Paper No. 2021/05.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack. 2022. "Selective Memory Equilibrium." https://doi.org/10.2139/ssrn.4015313.

Gabaix, Xavier. 2018. "Behavioral Inattention."

Gethin, Amory, Clara Martínez-Toledano, and Thomas Piketty. 2021. "Brahmin Left versus Merchant Right: Changing Political Cleavages in 21 Western Democracies, 1948-2020."

Ghaffari, Minou, and Susann Fiedler. 2018. "The Power of Attention: Using Eye Gaze to Predict Other-Regarding and Moral Choices." *Psychological Science*, 29(11): 1878–1889. https://doi.org/10.1177/0956797618799301.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*. https://doi.org/10.1086/701681.

Golman, Russell, George Loewenstein, Andras Molnar, and Silvia Saccardo. 2021. "The demand for, and avoidance of, information." *Management Science*.

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.

Gwinn, Rachael, Andrew B Leber, and Ian Krajbich. 2019. "The spillover effects of attentional learning on value-based choice." *Cognition*, 182: 294–306.

Gödker, Katrin, Peiran Jiao, and Paul Smeets. 2021. "Investor memory." *Available at SSRN 3348315*.

Hackett, M. J., and Nicholas. F. Gray. 2009. "Carbon Dioxide Emission Savings Potential of Household Water Use Reduction in the UK." *Journal of Sustainable Development*, 2(1): 36–43.

Heidhues, Paul, Botond Koszegi, and Philipp Strack. 2023. "Misinterpreting Yourself." *Available at SSRN 4325160*.

Hochleitner, Anna. 2022. "Fairness in times of crisis: Negative shocks, relative income and preferences for redistribution." CeDEx Discussion Paper Series.

Huffman, David, Collin Raymond, and Julia Shvets. 2022. "Persistent Overconfidence and Biased Memory: Evidence from Managers." *American Economic Review*, 112(10): 3141–3175. https://doi.org/10.1257/aer.20190668.

Hulshof, Daan, and Machiel Mulder. 2020. "Willingness to Pay for CO_2 Emission Reductions in Passenger Car Transport." *Environmental and Resource Economics*, 75(4): 899–929.

Hvidberg, Kristoffer B., Claus Kreiner, and Stefanie Stantcheva. 2020. "Social Position and Fairness Views." National Bureau of Economic Research.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. "Experimental designs for identifying causal mechanisms." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1): 5–51.

Imai, Taisuke, Davide D Pace, Peter Schwardmann, and Joël J van der Weele. 2022. "Correcting Consumer Misperceptions About CO 2 Emissions."

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475. https://doi.org/10.2307/2951620.

Jakiela, Pamela. 2015. "How fair shares compare: Experimental evidence from two cultures." *Journal of Economic Behavior & Organization*, 118: 40–54.

Jalil, Andrew J., Joshua Tasoff, and Arturo Vargas Bustamante. 2020. "Eating to Save the Planet: Evidence from a Randomized Controlled Trial Using Individual-Level Food Purchase Data." *Food Policy*, 95: 101950.

Jiang, Yuhong V, and Caitlin A Sisk. 2019. "Habit-like attention." *Current opinion in psychology*, 29: 65–70.

Judd, Charles M., and David A. Kenny. 1981. "Process analysis: Estimating mediation in treatment evaluations." *Evaluation Review*, 5(5): 602–619.

Just, Marcel A, and Patricia A Carpenter. 1980. "A theory of reading: from eye fixations to comprehension." *Psychological review*, 87(4): 329.

Kahana, Michael Jacob. 2012. *Foundations of human memory.* OUP USA.

Kallgren, Carl A, Raymond R Reno, and Robert B Cialdini. 2000. "A focus theory of normative conduct: When norms do and do not affect behavior." *Personality and Social Psychology Bulletin*, 26(8): 1002–1012.

Kanay, Ayşegül, Denis Hilton, Laetitia Charalambides, Jean-Baptiste Corrégé, Eva Inaudi, Laurent Waroquier, and Stéphane Cezera. 2021. "Making the Carbon Basket Count: Goal Setting Promotes Sustainable Consumption in a Simulated Online Supermarket." *Journal of Economic Psychology*, 83: 102348.

Kendall, Chad, and Ryan Oprea. 2021. "On the Complexity of Forming Mental Models." Working paper. 6.

Kendall, Chad W., and Constantin Charles. 2022. "Causal narratives." National Bureau of Economic Research.

Klenert, David, Linus Mattauch, Emmanuel Combet, Ottmar Edenhofer, Cameron Hepburn, Ryan Rafaty, and Nicholas Stern. 2018. "Making Carbon Pricing Work forCcitizens." *Nature Climate Change*, 8(8): 669–677.

Konovalov, Arkady, and Ian Krajbich. 2016. "Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning." *Nature communications*, 7(1): 1–11.

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

Koo, Hyunjin J., Paul K. Piff, and Azim F. Shariff. 2022. "If I Could Do It, So Can They: Among the Rich, Those With Humbler Origins are Less Sensitive to the Difficulties of the Poor." *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506221098921.

Kortelainen, Mika, Jibonayan Raychaudhuri, and Beatrice Roussillon. 2016. "Effects of Carbon Reduction Labels: Evidence from Scanner Data." *Economic Inquiry*, 54(2): 1167–1187.

Krajbich, Ian. 2019. "Accounting for attention in sequential sampling models of decision making." *Current opinion in psychology*, 29: 6–11.

Krajbich, Ian, Carrie Armel, and Antonio Rangel. 2010. "Visual fixations and the computation and comparison of value in simple choice." *Nature neuroscience*, 13(10): 1292.

Krajbich, Ian, Dingchao Lu, Colin Camerer, and Antonio Rangel. 2012. "The attentional drift-diffusion model extends to simple purchasing decisions." *Frontiers in psychology*, 3: 193.

Krawczyk, Michał. 2010. "A glimpse through the veil of ignorance: Equality of opportunity and support for redistribution." *Journal of Public Economics*, 94(1): 131–141. https://doi.org/10.1016/j.jpubeco.2009.10.003.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3): 495–524. https://doi.org/10.1111/jeea.12006.

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

Lefgren, Lars J., David P. Sims, and Olga B. Stoddard. 2016. "Effort, luck, and voting for redistribution." *Journal of Public Economics*, 143: 89–97. https://doi.org/10.1016/j.jpubeco.2016.08.012.

Le Yaouanq, Yves, and Peter Schwardmann. 2022. "Learning about one's self." *Journal of the European Economic Association*, 20(5): 1791–1828. Publisher: Oxford University Press.

Li, Shengwu. 2017. "Obviously strategy-proof mechanisms." *American Economic Review*, 107(11): 3257–87.

Lobeck, Max. 2021. "Motivating beliefs in a just world." Mimeo. 4 Working paper.

Lohmann, Paul, Elisabeth Gsottbauer, Anya Doherty, and Andreas Kontoleon. 2022. "Do Carbon Footprint Labels Promote Climatarian Diets? Evidence from a Large-Scale Field Experiment." *Journal of Environmental Economics and Management*, 114: 102693.

Löschel, Andreas, Bodo Sturm, and Carsten Vogt. 2013. "The Demand for Climate Protection–Empirical Evidence from Germany." *Economics Letters*, 118(3): 415–418.

Luo, M. Ronnier, Guihua Cui, and Bryan Rigg. 2001. "The development of the CIE 2000 colour-difference formula: CIEDE2000." *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5): 340–350. Publisher: Wiley Online Library.

Messick, David M, and Keith P Sentis. 1979. "Fairness and preference." *Journal of Experimental Social Psychology*, 15(4): 418–434.

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

Milosavljevic, Milica, Vidhya Navalpakkam, Christof Koch, and Antonio Rangel. 2012. "Relative visual saliency differences induce sizable bias in consumer choice." *Journal of Consumer Psychology*, 22(1): 67–74.

Mokrzycki, W. S., and Maciej Tatol. 2011. "Colour difference E-A survey." *Mach. Graph. Vis*, 20(4): 383–411.

Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality." *The Quarterly Journal of Economics*, 117(3): 735–774. https://doi.org/10.1162/003355302760193887.

Mullett, Timothy L, and Neil Stewart. 2016. "Implications of visual attention phenomena for models of preferential choice." *Decision*, 3(4): 231.

Müller, Maximilian W. 2022. "Selective Memory around Big Life Decisions." Working papers.

Nagel, Rosemarie. 1995. "Unraveling in guessing games: An experimental study." *The American economic review*, 85(5): 1313–1326. Publisher: JSTOR.

Nemet, Gregory F., and Evan Johnson. 2010. "Willingness to Pay for Climate Policy: A Review of Estimates." La Follette School Working Paper No. 2010-011.

Newell, Richard G., and Juha Siikamäki. 2014. "Nudging Energy Efficiency Behavior: The Role of Information Labels." *Journal of the Association of Environmental and Resource Economists*, 1(4): 555–598.

Oprea, Ryan. 2022. "Simplicity equivalents." Working Paper.

Orquin, Jacob L., and Simone Mueller Loose. 2013. "Attention and choice: A review on eye movements in decision making." *Acta Psychologica*, 144(1): 190–206. https://doi.org/10.1016/j.actpsy.2013.06.003.

Ostling, Robert. 2009. "Economic Influences on Moral Values." *The B.E. Journal of Economic Analysis & Policy*, 9(1). https://doi.org/10.2202/1935-1682.2044.

Pace, Davide D., and Joël van der Weele. 2020. "Curbing Carbon: An Experiment on Uncertainty and Information about CO_2 Emissions." Tinbergen Institute Discussion Paper No. 2020-059.

Pachur, Thorsten, Michael Schulte-Mecklenbeck, Ryan O. Murphy, and Ralph Hertwig. 2018. "Prospect theory reflects selective allocation of attention." *Journal of Experimental Psychology: General*, 147(2): 147–169. https://doi.org/10.1037/xge0000406.

Padgett, J. Paul, Anne C. Steinemann, James H. Clarke, and Michael P. Vandenbergh. 2008. "A Comparison of Carbon Calculators." *Environmental Impact Assessment Review*, 28(2-3): 106–115.

Perino, Grischa, Luca A. Panzone, and Timothy Swanson. 2014. "Motivation Crowding in Real Consumption Decisions: Who Is Messing with My Groceries?" *Economic Inquiry*, 52(2): 592–607.

Piff, Paul K., Dylan Wiwad, Angela R. Robinson, Lara B. Aknin, Brett Mercier, and Azim Shariff. 2020. "Shifting attributions for poverty motivates opposition to inequality and enhances egalitarianism." *Nature Human Behaviour*, 4(5): 496–505. https://doi.org/10.1038/s41562-020-0835-8.

Piketty, Thomas. 2020. *Capital and ideology.* Harvard University Press.

Poore, Joseph, and Thomas Nemecek. 2018. "Reducing Food's Environmental Impacts through Producers and Consumers." *Science*, 360(6392): 987–992.

Potter, Mary C. 1976. "Short-term conceptual memory for pictures." *Journal of experimental psychology: human learning and memory*, 2(5): 509.

Potter, Mary C, Brad Wyble, Carl Erick Hagmann, and Emily S McCourt. 2014. "Detecting meaning in RSVP at 13 ms per picture." *Attention, Perception, & Psychophysics*, 76(2): 270–279.

Pärnamets, Philip, Petter Johansson, Lars Hall, Christian Balkenius, Michael J. Spivey, and Daniel C. Richardson. 2015. "Biasing moral decisions by exploiting the dynamics of eye gaze." *Proceedings of the National Academy of Sciences*, 112(13): 4170–4175.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292. https://www.jstor.org/stable/2999450 (accessed 2023-04-21). Publisher: [Wiley, Econometric Society].

Rabin, Matthew, and Georg Weizsäcker. 2009. "Narrow Bracketing and Dominated Choices." *American Economic Review*, 99(4): 1508–1543.

Rahal, Rima-Maria, and Susann Fiedler. 2019. "Understanding cognitive and affective mechanisms in social psychology through eye-tracking." *Journal of Experimental Social Psychology*, 85: 103842.

Rahal, Rima-Maria, Susann Fiedler, and Carsten KW De Dreu. 2020. "Prosocial preferences condition decision effort and ingroup biased generosity in intergroup decision-making." *Scientific reports*, 10(1): 1–11.

Rodemeier, Matthias, and Andreas Löschel. 2020. "The Welfare Effects of Persuasion and Taxation: Theory and Evidence from the Field." Available at SSRN: http://dx.doi.org/10.2139/ssrn.3587339.

Rodriguez-Lara, Ismael, and Luis Moreno-Garrido. 2012. "Self-interest and fairness: self-serving choices of justice principles." *Experimental Economics*, 15(1): 158–175.

Rustichini, Aldo, and Marie Claire Villeval. 2014. "Moral hypocrisy, power and social preferences." *Journal of Economic Behavior & Organization*, 107: 10–24.

Saccardo, Silvia, and Marta Serra-Garcia. 2023. "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment." *American Economic Review*, 113(2): 396–429.

Salazar, Miguel, Daniel Joel Shaw, Kristína Czekóová, Rostislav Staněk, and Milan Brázdil. 2022. "The role of generalised reciprocity and reciprocal tendencies in the emergence of cooperative group norms." *Journal of Economic Psychology*, 90: 102520.

Sandel, Michael J. 2020. *The Tyranny of Merit: What's become of the common good?* Allen Lane London.

Saucet, Charlotte, and Marie Claire Villeval. 2019. "Motivated memory in dictator games." *Games and Economic Behavior*, 117: 250–275. https://doi.org/10.1016/j.geb.2019.05.011.

Schlag, Karl H., and Joël van der Weele. 2013. "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality." *Theoretical Economics Letters*, 3: 38–42.

Schlag, Karl H., James Tremewan, and Joël J. van der Weele. 2015. "A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs." *Experimental Economics*, 18(3): 457–490.

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and self-deception." *Nature human behaviour*, 3(10): 1055–1061. Publisher: Nature Publishing Group UK London.

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

Shi, Jing, Vivianne H. M. Visschers, Michael Siegrist, and Joseph Arvai. 2016. "Knowledge as a Driver of Public Perceptions about Climate Change Reassessed." *Nature Climate Change*, 6(8): 759–762.

Shimojo, Shinsuke, Claudiu Simion, Eiko Shimojo, and Christian Scheier. 2003. "Gaze bias both reflects and influences preference." *Nature Neuroscience*, 6(12): 1317–1322. `https://doi.org/10.1038/nn1150`.

Smith, Anna Jo Bodurtha, Imogen Tennison, Ian Roberts, John Cairns, and Caroline Free. 2013. "The Carbon Footprint of Behavioural Support Services for Smoking Cessation." *Tobacco Control*, 22(5): 302–307.

Smith, Stephanie M, and Ian Krajbich. 2018. "Attention and choice across domains." *Journal of Experimental Psychology: General*, 147(12): 1810.

Smith, Stephanie M, and Ian Krajbich. 2019. "Gaze amplifies value in decision making." *Psychological science*, 30(1): 116–128.

Soregaroli, Claudio, Elena Claire Ricci, Stefanella Stranieri, Rodolfo M. Nayga Jr, Ettore Capri, and Elena Castellari. 2021. "Carbon Footprint Information, Prices, and Restaurant Wine Choices by Customers: A Natural Field Experiment." *Ecological Economics*, 186: 107061.

Spaargaren, Gert, C. S. A. Van Koppen, Anke M. Janssen, Astrid Hendriksen, and Corine J. Kolfschoten. 2013. "Consumer Responses to the Carbon Labelling of Food: A Real Life Experiment in a Canteen Practice." *Sociologia Ruralis*, 53(4): 432–453.

Sparkman, Gregg, Nathan Geiger, and Elke U Weber. 2022. "Americans experience a false social reality by underestimating popular climate policy support by nearly half." *Nature Communications*, 13(1): 1–9.

Stango, Victor, and Jonathan Zinman. 2020. "We are all behavioral, more or less: A taxonomy of consumer decision making." National Bureau of Economic Research.

Suhay, Elizabeth, Marko Klasnja, and Gonzalo Rivero. 2020. "Ideology of Affluence: Rich Americans' Explanations for Inequality and Attitudes toward Redistribution."

Suhay, Elizabeth, Marko Klašnja, and Gonzalo Rivero. 2021. "Ideology of affluence: Explanations for inequality and economic policy preferences among rich Americans." *The Journal of Politics*, 83(1): 367–380.

Taufique, Khan M. R., Kristian S. Nielsen, Thomas Dietz, Rachael Shwom, Paul C. Stern, and Michael P. Vandenbergh. 2022. "Revisiting the Promise of Carbon Labelling." *Nature Climate Change*, 12(2): 132–140.

Theeuwes, Jan. 2019. "Goal-driven, stimulus-driven, and history-driven selection." *Current opinion in psychology*, 29: 97–101.

Thøgersen, John. 2008. "Social norms and cooperation in real-life social dilemmas." *Journal of Economic Psychology*, 29(4): 458–472.

Tobler, Christina, Vivianne H. M. Visschers, and Michael Siegrist. 2012. "Consumers' Knowledge About Climate Change." *Climatic Change*, 114(2): 189–209.

Toussaert, Séverine. 2018. "Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment." *Econometrica*, 86(3): 859–889. https://doi.org/10.3982/ECTA14172.

Ubeda, Paloma. 2014. "The consistency of fairness rules: An experimental study." *Journal of Economic Psychology*, 41: 88–100. https://doi.org/10.1016/j.joep.2012.12.007.

UK BEIS. 2020. "Greenhouse Gas Reporting: Conversion Factors 2019." UK Department for Business, Energy & Industrial Strategy. https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019.

Valero, Vanessa. 2021. "Redistribution and beliefs about the source of income inequality." *Experimental Economics*. https://doi.org/10.1007/s10683-021-09733-8.

VanderWeele, Tyler, and Stijn Vansteelandt. 2014. "Mediation analysis with multiple mediators." *Epidemiologic methods*, 2(1): 95–115. Publisher: De Gruyter.

Vetter, David. 2020. "Got Beef? Here's What Your Hamburger Is Doing to the Climate." *Forbes*, October 5. https://www.forbes.com/sites/davidrvetter/2020/10/05/got-beef-heres-what-your-hamburger-is-doing-to-the-climate.

Visschers, Vivianne H. M., and Michael Siegrist. 2015. "Does Better for the Environment Mean Less Tasty? Offering More Climate-Friendly Meals Is Good for the Environment and Customer Satisfaction." *Appetite*, 95: 475–483.

Vlaeminck, Pieter, Ting Jiang, and Liesbet Vranken. 2014. "Food Labeling and Eco-Friendly Consumption: Experimental Evidence from a Belgian Supermarket." *Ecological Economics*, 108: 180–190.

Waldfogel, Hannah B., Jennifer Sheehy-Skeffington, Oliver P. Hauser, Arnold K. Ho, and Nour S. Kteily. 2021. "Ideology selectively shapes attention to inequality." *Proceedings of the National Academy of Sciences*, 118(14). https://doi.org/10.1073/pnas.2023985118.

Waxman, Henry A., and Edward J. Markey. 2009. "American Clean Energy and Security Act of 2009." Washington: US House of Representatives. https://www.congress.gov/bill/111th-congress/house-bill/2454.

Willemsen, Martijn C., and Eric J. Johnson. 2019. "Observing Cognition with MouselabWEB." In *A handbook of process tracing methods*. 76–95.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs." *American Economic Review*, 110(2): 337–361. https://doi.org/10.1257/aer.20180728.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. For a full list of PhD theses that appeared in the series we refer to List of PhD Theses – Tinbergen.nl. The following books recently appeared in the Tinbergen Institute Research Series:

775 A. CASTELEIN, *Models for Individual Responses*

776 D. KOLESNYK, *Consumer Disclosures on Social Media Platforms: A Global Investigation*

777 M.A. ROLA-JANICKA, *Essays on Financial Instability and Political Economy of Regulation*

778 J.J. KLINGEN, *Natural Experiments in Environmental and Transport Economics*

779 E.M. ARTMANN, *Educational Choices and Family Outcomes*

780 F.J. OSTERMEIJER, *Economic Analyses of Cars in the City*

781 T. ÖZDEN, *Adaptive Learning and Monetary Policy in DSGE Models*

782 D. WANG, *Empirical Studies in Financial Stability and Natural Capital*

783 L.S. STEPHAN, *Estimating Diffusion and Adoption Parameters in Networks New Estimation Approaches for the Latent-Diffusion-Observed-Adoption Model*

784 S.R. MAYER, *Essays in Financial Economics*

785 A.R.S. WOERNER, *Behavioral and Financial Change – Essays in Market Design*

786 M. WIEGAND, *Essays in Development Economics*

787 L.M. TREUREN, *Essays in Industrial Economics - Labor market imperfections, cartel stability, and public interest cartels*

788 D.K. BRANDS, *Economic Policies and Mobility Behaviour*

789 H.T.T. NGUYỄN, *Words Matter? Gender Disparities in Speeches, Evaluation and Competitive Performance*

790 C.A.P BURIK *The Genetic Lottery. Essays on Genetics, Income, and Inequality*

791 S.W.J. OLIJSLAGERS, *The Economics of Climate Change: on the Role of Risk and Preferences*

792 C.W.A. VAN DER KRAATS, *On Inequalities in Well-Being and Human Capital Formation*

793 Y. YUE, *Essays on Risk and Econometrics*

794 E.F. JANSSENS, *Estimation and Identification in Macroeconomic Models with Incomplete Markets*

795 P.B. KASTELEIN, *Essays in Household Finance: Pension Funding, Housing and Consumption over the Life Cycle*

796 J.O. OORSCHOT, *Extremes in Statistics and Econometrics*

797 S.D.T. HOEY, *Economics on Ice: Research on Peer Effects, Rehiring Decisions and Worker Absenteeism*

798 J. VIDIELLA-MARTIN, *Levelling the Playing Field: Inequalities in early life conditions and policy responses*

799 Y. XIAO, *Fertility, parental investments and intergenerational mobility*

800 X. YU, *Decision Making under Different Circumstances: Uncertainty, Urgency, and Health Threat*

801 G. GIANLUCA, *Productivity and Strategies of Multiproduct Firms*

802 H. KWEON, *Biological Perspective of Socioeconomic Inequality*

803 D.K. DIMITROV, *Three Essays on the Optimal Allocation of Risk with Illiquidity, Intergenerational Sharing and Systemic Institutions*

804 J.B. BLOOMFIELD, *Essays on Early Childhood Interventions*

805 S. YU, *Trading and Clearing in Fast-Paced Markets*

806 M.G. GREGORI, *Advanced Measurement and Sampling for Marketing Research*

807 O.C. SOONS, *The Past, Present, and Future of the Euro Area*

808 D. GARCES URZAINQUI *The Distribution of Development. Essays on Economic Mobility, Inequality and Social Change*

809 A.C. PEKER, *Guess What I Think: Essays on the Wisdom in Meta-predictions*

810 A. AKDENIZ, *On the Origins of Human Sociality*

811 K. BRÜTT, *Strategic Interaction and Social Information: Essays in Behavioural Economics*

812 P.N. KUSUMAWARDHANI, *Learning Trends and Supply-side Education Policies in Indonesia*

813 F. CAPOZZA, *Essays on the Behavioral Economics of Social Inequalities*

814 D.A. MUSLIMOVA, *Complementarities in Human Capital Production: The Role of Gene-Environment Interactions*

815 J.A. DE JONG, *Coordination in Market and Bank Run Experiments*

816 Y. KIM, *Micro studies of macroprudential policies using loan-level data*

817 S.R. TER MEULEN, *Grade retention, ability tracking, and selection in education*

818 A.G.B. ZIEGLER, *The Strategic Role of Information in Markets and Games: Essays in Behavioral Economics*

819 I. VAN DER WERVE, *Panel data model for socioeconomic studies in crime and education*

820 Y. GU, *Roads, Institutions and the Primary Sector in West Africa*

821 Y. LI, *Share Repurchases in the US: An extensive study on the data, drivers, and consequences*

822 R. DIAS PEREIRA, *What Makes us Unique? Genetic and Environmental Drivers of Health and Education Inequalities*

823 H.P. LETTERIE, *Essays on the regulation of long-term care in the Netherlands*