

BEYOND BAYES: INFORMATION, BELIEFS, AND ECONOMIC OUTCOMES

MARCELO WOO

PhD advisors: Silvia Sonderegger & Martin Sefton



*Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy*

SCHOOL OF ECONOMICS
UNIVERSITY OF NOTTINGHAM

September 2025

ACKNOWLEDGEMENTS

This thesis is the product of countless influences that people around me have had on me. Writing it would not have been possible without all the incredible people I met prior and during the PhD, and I feel deep gratitude and appreciation for everyone who accompanied and supported me throughout this special journey.

First, I would like to thank my supervisors, Silvia Sonderegger and Martin Sefton, and my former supervisor, Fabio Tufano, for their support and encouragement. Their mentorship, expertise, and guidance have shaped not only this thesis but also my journey as a researcher. My gratitude extends to Simon Gächter for providing very valuable feedback in several instances.

I also want to thank Christian Ruzzier, co-author on one of my research projects. Working together was a highly instructive experience and I learned a lot from his way of thinking and the discussions we had both during the project and outside of it.

Ultimately, the experience of my PhD would not have been the same without the stimulating environment in the School of Economics, and in particular the Centre for Decision Research and Experimental Economics (CeDEx). Participating and presenting in seminars had an immense impact on me and I am grateful for all the support and feedback I received over the last years.

This PhD has been possible thanks to several sources of funding. I am very grateful for the scholarships provided by the School of Economics, the Research Institute for Development, Growth, and Economics (RIDGE), and my generous sponsor Richard Owen, and for the financial support for my experimental research provided by CeDEx and Leicester University.

I would also like to express my sincere gratitude to Florian Zimmermann for hosting me during my research visit at IZA and the University of Bonn. This experience was deeply enriching — both personally and academically — and gave me the opportunity to engage in stimulating discussions and connect with many wonderful people.

In addition to all the scientific support I received, I am also incredibly grateful for the community of fellow PhD students that I experienced in Nottingham and all the friends I made along the way. In particular, I want to thank Matías Golman, Jesús Rodríguez, Guido Lamarmora, Agustina Sampaolesi, Nicolás Pazos, Fabrizio Ardilles, Lucas de Melo, Anna Hochleitner, Malte Baader, Sarah Bowen, Candice Ford, Richard Mills, Jose Guinot Saporta, Lara Suraci, Yuliet Verbel Bustamante, and Monika Pompeo. To my dear and loyal friends beyond the PhD: Daniel and Samuel Cho, Sofía Vucinovich, Martin Kim, Adriana An, Diego Lee, Elias and Elena Lee, Martin (Dae Jung) Kim, Paola Jeon, Marcos Fabietti, Petru Gânj, Mohammed Musa, Iain Messori, Tomás Cedeño, Hasan Syed, Artem Prokhorov, Andrés Sandoval, Elizabeth Brown, Mariana Cabral, Mariana Gómez Gil, Maren Wieler, and Fotini Iacovou. Thanks for the pleasure of creating so many great memories, laughs, and adventures together during this chapter of my life. My PhD experience would have never been the same without you.

Last, but not least, I am deeply grateful to my family: Sang Bum, Jong Hi, Jonathan, Angie, and Justo for your constant support, love, and encouragement. Your presence, even from afar, has been a source of strength and motivation and made the journey more enjoyable.

CONTENTS

GENERAL INTRODUCTION	1
I BURYING HEADS IN THE SAND TOGETHER: THEORY AND EXPERIMENT ON THE INTERDEPENDENCE OF INFORMATION AVOIDANCE	4
1 Introduction	4
2 Related literature	7
3 Theoretical framework	8
3.1 Types of best response (“Strategy types”)	10
4 Experimental design and procedures	12
4.1 Experimental design	12
4.1.1 Decision environment	12
4.1.2 Treatments	12
4.1.3 Individual best-response schedules and heterogeneity	13
4.1.4 Additional measurements	13
4.1.5 Task and earnings	15
4.2 Experimental procedures	15
5 Results	19
5.1 Preliminary analyses	20
5.1.1 Are screams a bad?	20
5.1.2 How prevalent is information avoidance in groups?	20
5.2 The interdependence of information avoidance: experimental evidence	22
5.3 Heterogeneity of interdependence	28
5.4 Mechanisms	31
5.4.1 Anticipatory utility and the Mutually Assured Ignorance mechanism	32
5.4.2 Other mechanisms	36
6 Group composition and equilibria	38
7 Conclusion	46
II CONFIRMATION-BIASED EVALUATION AND LABOR MARKET DISCRIMINATION	51
1 Introduction	51
2 A model of learning with confirmation bias	55
3 Discrimination in the labor market	58
4 Discrimination and segregation	61
5 Discrimination and competition	65
6 Concluding remarks	66

APPENDICES	75
A Appendix for Chapter 1	75
A.1 Theoretical appendix	75
A.1.1 Proofs	75
A.2 Additional analyses	77
A.2.1 Descriptive statistics	77
A.2.2 Are screams a bad? Preferences over screams and volume	77
A.2.3 Balance checks	79
A.2.4 Robustness to exclusion of scream- and volume-lovers	81
A.2.5 Robustness to alternative specifications	84
A.2.6 Alternative outcome: continuous measure of avoidance (IA Scale)	87
A.2.7 Instrumental variable regressions (2SLS)	89
A.2.8 Mechanisms — supplementary tables	91
A.3 Instructions and other screens	92
B Appendix for Chapter 2	117
B.1 Dynamics of confirmation-biased learning	117

LIST OF FIGURES

Figure 1.1	Examples of best-response functions, by strategy type	14
Figure 1.2	Timeline of experiment	15
Figure 1.3	Information avoidance: incidence and preference strength	21
Figure 1.4	Manipulation check	23
Figure 1.5	Information avoidance: treatment effect and predictors	24
Figure 1.6	Heterogeneity of strategy types	28
Figure 1.7	Switching thresholds and beliefs by strategy type and treatment condition	31
Figure 1.8	Association between net expected anticipatory utility and information avoidance	35
Figure 1.9	Mechanisms: treatment effects on components of net anticipatory utility	37
Figure 1.10	Aggregate best-response functions and information equilibria (homogeneous groups)	40
Figure 1.11	Aggregate best-response function and information equilibria (full empirical sample)	41
Figure 1.12	Effect of group composition on aggregate best-response schedules	42
Figure 1.13	Group composition and the distribution of equilibrium shares of avoiders	44
Figure 1.14	Effect of group composition on distribution of information equilibria	45
Figure 1.15	Group composition and features of equilibrium distributions	47
Figure A.1	Preferences over volume of screams: reported utilities and preference types	78
Figure A.2	Preferences over screams: reported utilities upon learning each state	79
Figure A.3	Treatment effect on information avoidance by preference over screams and volume	81
Figure A.4	Specification robustness plot ($\alpha = 0.05$)	85
Figure A.5	Specification robustness plot ($\alpha = 0.01$)	86

LIST OF TABLES

Table 1.1	Manipulation check (regressions)	25
Table 1.2	Treatment effect on information avoidance	27
Table 1.3	Treatment effect on distribution of strategy types (extensive margin)	29
Table 1.4	Treatment effect on information avoidance, by strategy type (intensive margin)	30
Table 1.5	Anticipatory utility and information avoidance	34
Table A.1	Sample summary statistics	77
Table A.2	Distribution of Scream- and Volume-preference types	78
Table A.3	Balance checks	80
Table A.4	Treatment effect on information avoidance (subsample of scream haters)	82
Table A.5	Treatment effect on information avoidance (subsample of volume haters)	83
Table A.6	Post-double selection estimates	85
Table A.7	Treatment effect on IA Scale	88
Table A.8	Effect of beliefs on information avoidance	90
Table A.9	Interaction of beliefs and information avoidance	91

GENERAL INTRODUCTION

Information and beliefs are central inputs to economic decision-making. From individual choices in education, finance, and consumption to collective decisions in organizations, politics, and markets, economic outcomes depend on how individuals acquire and process information, update their beliefs, and act upon them. In the standard economic framework, information is treated as a typically valuable and desirable resource: when available at no cost, rational agents are predicted to acquire it to improve the quality of their decisions and associated payoffs (Stigler, 1961). As the theory goes, once information is obtained, agents process it objectively much like a cold-headed statistician would, updating beliefs according to Bayes' rule, with the implication that, as the number of signals grows, inaccurate beliefs are eventually corrected (DeGroot, 1970).

In practice, however, attitudes toward information and belief updating processes often depart from this benchmark. A growing body of evidence shows that, in some cases, individuals deliberately avoid exposure to decision-relevant information, even when it is freely available to them (Thornton, 2008; Karlsson et al., 2009; Oster et al., 2013; Ganguly & Tasoff, 2017; Golman et al., 2017; Hertwig & Engel, 2020). In other cases, conditional on observing information, they (mis)interpret the evidence in ways that reinforce pre-existing views (Lord et al., 1979; Nickerson, 1998; Rabin & Schrag, 1999) or allow them to maintain favorable and good-feeling beliefs about themselves and the world (Eil & Rao, 2011; Bénabou, 2015; Bénabou & Tirole, 2016; Engelmann et al., 2024). Both types of behavior (ex-ante avoidance and ex-post signal distortion) depart from the standard Bayesian model of learning. While often studied at the individual level, such behaviors can nonetheless generate systematic distortions in collective outcomes, with implications for welfare, efficiency, and inequality. Understanding how such non-standard learning affects aggregate outcomes is therefore essential for advancing economic theory and informing policy design.

This thesis contributes to this line of research by examining two settings in which agents may deviate from Bayesian learning: (i) the avoidance of instrumentally valuable, but potentially discomfort-inducing information in groups and organizations, which can degrade the quality of decisions, with potentially large collective consequences, and (ii) psychologically biased evaluation of signals of workers' productivity leading to persistent discrimination and segregation in labor markets. The title *Beyond Bayes: Information, Beliefs, and Economic Outcomes* reflects the common theme of the thesis: each chapter studies how well-documented individual-level departures from the Bayesian model of information processing and belief updating can affect aggregate economic outcomes.

The first chapter investigates the phenomenon of **information avoidance in groups**. Most existing studies document and analyze this behavior by individuals in isolation (see, e.g., Kreps & Porteus, 1978; Dana et al., 2007; Carrillo & Mariotti, 2000; Golman et al., 2017). Yet many real-world settings (such as climate change mitigation, financial bubbles, and the undertaking of risky projects by organizations) involve interdependent payoffs, where one person's decision to "bury his head in the sand" can directly impact others' material welfare.

Accordingly, Chapter I considers settings where avoidance of information imposes *negative payoff externalities* and asks whether such decisions **can be interdependent**: does an individual's propensity to remain ignorant increase (or decrease) in her expectation that others will do the same? To address this question, the chapter develops a theoretical framework building on Bénabou (2013) and, informed by the model, designs a novel laboratory experiment to test whether information avoidance can be interdependent. The results show that, on average, individuals are more likely to avoid information when they expect others to do the same, creating the possibility of collectively ignorant equilibria. At the same time, we uncover substantial heterogeneity beyond this average tendency: some individuals' responses exhibit strategic complementarity, others' substitutability, while some others' are unconditional. As we show through simulations, this diversity in "strategy types" implies that group composition plays a crucial role in determining equilibrium levels of avoidance. By documenting such interdependences in information decisions, the chapter highlights the existence of "cognitive externalities" —where one person's information decision alters the psychological incentives of others— which carry important implications for organizational design and for policies aimed at increasing overall acquisition of information.

The second chapter moves away from the behavior of ex-ante avoidance of information, and looks at the consequences of ex-post distortion of signals: it turns to **confirmation bias** in belief updating as **a source of persistent discrimination** in the labor market. Based on joint work with Christian Ruzzier, the chapter builds on the literature on inaccurate statistical discrimination (Bohren et al., 2025) and shows how employers who start with incorrect stereotypes about the productivity of minority groups and who are confirmation-biased —so that they interpret ambiguous signals in a way that aligns with their priors— fail to fully correct their prior beliefs, even after observing a large number of worker performance signals. The model predicts persistent wage gaps across social groups (e.g., gender, race) and patterns of segregation, and provides insights into why standard policy interventions may be less effective when confirmation bias affects employers' learning.

While the two chapters examine different facets of non-Bayesian learning (ex-ante avoidance and ex-post signal distortion), this thesis shows how individual-level departures in information processing and belief updating can generate aggregate economic outcomes with broader organizational and societal consequences.

References

- Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *The Review of Economic Studies*, 80(2), 429–462.
- Bénabou, R. (2015). The Economics of Motivated Beliefs:. *Revue d'économie politique*, Vol. 125(5), 665–685.
- Bénabou, R. & Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, 30(3), 141–164.
- Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2025). Inaccurate Statistical Discrimination: An Identification Problem. *The Review of Economics and Statistics*, 107(3), 605–620.
- Carrillo, J. D. & Mariotti, T. (2000). Strategic Ignorance as a Self-Disciplining Device. *The Review of Economic Studies*, 67(3), 529–544.
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
- Eil, D. & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, 3(2), 114–138.
- Engelmann, J. B., Lebreton, M., Salem-Garcia, N. A., Schwardmann, P., & van der Weele, J. J. (2024). Anticipatory Anxiety and Wishful Thinking. *American Economic Review*, 114(4), 926–960.
- Hertwig, R. & Engel, C., Eds. (2020). *Deliberate Ignorance: Choosing Not to Know*. Strüngmann Forum Reports. Cambridge, Massachusetts: The MIT Press.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2), 175–220.
- Rabin, M. & Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, 114(1), 37–82.
- Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98(5), 1829–1863.

CHAPTER 1

Burying heads in the sand together: theory and experiment on the interdependence of information avoidance[†]

1 Introduction

A central view in economics is that information is a valuable resource: by acquiring it, agents can typically improve their decisions and attain higher payoffs (Stigler, 1961). Agents are thus predicted to welcome information about unknown states of the world. In contrast with this prediction, a growing body of evidence documents the opposite behavior: individuals often engage in “information avoidance” —the deliberate refusal of decision-relevant information that is readily available to obtain and use— (Golman et al., 2017; Hertwig & Engel, 2020), rather acting as if “ignorance is bliss”. For instance, research has shown that people refuse medical tests to avoid learning about their health conditions, reduce exposure to news relevant to their own investments, pay not to learn how they rank in intelligence and beauty, shun opposing political or religious views, and disengage from news about climate change and other collective risks (Thornton, 2008; Karlsson et al., 2009; Eil & Rao, 2011; Oster et al., 2013; Ganguly & Tasoff, 2017).

While information avoidance has been studied at the individual level, much less is known about how it interacts and spreads within groups. In many collective settings (from mitigating climate change and infectious diseases to ensuring a company’s financial health), individual outcomes are interlinked: one person’s decision to remain ignorant about the state of the world can impose *negative externalities* on others by reducing her ability to act effectively (e.g., resulting in larger climate risks, increased probability of infections, and higher likelihood of bankruptcy and unemployment for everyone). As a result, the *prospects* faced by each individual are *endogenous* to the information decisions of other members. As Bénabou (2013) shows, if agents derive utility from their beliefs about the state of the world, this endogeneity may create a vicious cycle: as more agents avoid news about the state, the future worsens, making the

[†]This chapter is based on joint work with Silvia Sonderegger and Fabio Tufano. I am grateful to CeDEx and to the University of Leicester for funding this research. I also thank CeDEx members and participants at the CCC conference, IZA’s Applied Lunch seminar, and LISER Field Days conference for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 28/11/2023.

truth even harder to confront. Information decisions (to acquire or avoid information) can thus become strategic complements. In the most extreme cases, this complementarity may trap the group in an equilibrium of collective blindness to information, where individuals overlook looming crises and fail to coordinate on their mitigation, despite signals being readily available. History offers examples of organizations shunning information away and becoming “color-blind in a sea of red flags”, with catastrophic consequences.¹

This theoretical possibility raises a couple of key questions. First, can the interdependence of information avoidance arise empirically? Put differently, can willful ignorance be contagious, spreading within groups as individuals anticipate worse prospects due to others remaining uninformed? Second, can there be other forms of interdependence beyond contagion? If so, how large is the heterogeneity?

This paper answers these questions by combining theory, experiment, and simulations. Our investigation begins with a theoretical model inspired by [Bénabou \(2013\)](#), which studies the strategic interdependencies that may arise in group settings. As well as contagious ignorance, we show that in our setup there is also an alternative possibility. Depending on the nature of their utility function, individuals may exhibit strategic substitutability in information avoidance: when others bury their heads in the sand, these individuals become less inclined to shun information. Furthermore, the theory also shows that some individuals may not base their choices on the actions of others at all, and may instead unconditionally avoid or unconditionally acquire information, regardless of what others do. These findings indicate that the dynamics of information avoidance in groups may vary considerably, a point we explore with simulations based on empirical data.

Informed by this theory, we design a novel lab experiment to assess the interdependence of information avoidance empirically. Participants are matched in groups and informed that there are two possible states of the world: *Good* and *Bad*. In the *Bad* state, everyone in the group will be exposed to an aversive future event: at the end of the experiment, they will hear a series of distressing screams through headphones. In the *Good* state, no noises will be heard. At the beginning of the experiment, each participant chooses whether to discover the state immediately (option “Now”) or to wait until the future event unfolds (option “Later”). Importantly, the volume of the screams in the *Bad* state decreases in the number of members who acquire information about the state (that is, who choose Now). In other words, each participant understands that by selecting Now, they can mitigate the intensity of the noises not only for themselves but also for everyone else in their group, should the state turn out to be *Bad*.

In order to assess the interdependence, we are interested in how information decisions respond to different beliefs about how others deal with information. We employ two empirical strategies for this purpose. First, we exogenously vary subjects’ expectations about the share of avoiders in their group by informing them about the analogous share observed in a pilot study. Two treatments are randomized at the level of experimental groups: in the *Many Avoided*

¹For instance, Merrill Lynch during the 2008 financial crisis, General Motors prior to its 2009 collapse, and NASA in the Columbia Shuttle disaster illustrate organizational failures driven by aversion to uncomfortable information.

condition, subjects are told that 80% of subjects in a prior group avoided information, whereas in the *Few Avoided* condition, subjects are told only 20% did. The resulting variation in posterior beliefs allows us to test whether, on average, a higher expected avoidance among peers increases own avoidance. Second, we use the strategy method (Selten (1967); Fischbacher et al. (2001)) to elicit, for each individual, conditional responses to varying shares of avoiders within the group, allowing us to identify individual heterogeneity in best-response schedules beyond the average effect.

We find that our treatment intervention strongly shifts participants' expectations about their peers: beliefs about the share of avoiders increase by 22 percentage points on average, from 37% in *Few Avoided* to 59% in *Many Avoided*. The treatment also affects behavior: consistent with contagious ignorance, the share of subjects avoiding information increases by one-third, from 24% to 33%. To explore mechanisms, we elicit expected anticipatory utilities prior to the information decision. In line with the theoretical channel, subjects in the *Many Avoided* condition report lower anticipatory utility (relative to those in *Few Avoided*) when imagining the event of acquiring information and then finding out *bad* news about the state—plausibly because higher expected avoidance rates increase the severity of the noises in the bad state—, but no difference for the event of finding out *good* news.

While the average effects point to contagion, individual responses vary considerably: in line with the model, we document substantial heterogeneity in how participants respond to varying shares of information avoiders in their group. Among those influenced by others' choices (about 39% of the sample), half exhibit a pattern consistent with contagious ignorance (strategic complementarity), while the other half display the opposite tendency (strategic substitutability). Hence, although many participants exhibit preferences consistent with contagious ignorance, an equally sizable share becomes more willing to acquire information when they expect widespread avoidance.

The diversity of such strategy types underscores the importance of group composition for the equilibrium prevalence of avoidance. Motivated by this observation, we employ simulations to show how the distribution of strategy types determines the aggregate level of information avoidance. Groups primarily composed of individuals who exhibit strategic complementarity in information acquisition are at risk of getting trapped in a so-called *Mutually Assured Ignorance* (MAI) equilibrium (see Section 3), where no one acquires information. The inclusion of unconditional avoiders precipitates this outcome: with 30% unconditional avoiders, the MAI equilibrium arises with 75% probability, and with 50% it emerges with certainty. Likewise, unconditional information getters can drive the group toward a virtuous equilibrium where everybody acquires information (*Mutually Assured Awareness* equilibrium). In contrast, groups composed of individuals who exhibit strategic substitutability never reach a MAI equilibrium but are also unable to sustain an equilibrium characterized by universal information acquisition. In addition, our simulations show that the relationship between group composition and the equilibrium incidence of information avoidance is nonlinear: an identical amount of change (e.g., a 10 percentage point increase in the participation of unconditional avoiders) can have very different effects on the equilibrium share of avoiders, depending on the initial distribution of types.

Our findings suggest implications for policy-makers and organization designers (“principals”) seeking to increase information acquisition rates in their organizations and populations of interest. While most of the literature on information preferences examines individuals in isolation, both our model and empirical results show that information decisions can be interdependent, in such a way that incentivizing an individual to acquire information may encourage—or dissuade—others from doing so. In other words, information choices may generate “cognitive externalities” that amplify or offset the effectiveness of such incentives. Optimal incentive schemes should therefore account for such “crowding in” and “crowding out” effects in a general equilibrium framework. A second policy lever is available when the principal assembles teams in organizations. Our simulations show that group composition by strategy type can give rise to widely different equilibria, suggesting that principals may benefit from considering these types and their interactions when designing teams.

The remainder of the paper is organized as follows. In Section 2, we discuss related literature. In Section 3, we lay out our theoretical framework. In Section 4, we present the experimental design. In Section 5, we report the empirical findings. In Section 6, we use our data to inform the relationship between group compositions and equilibrium levels of avoidance. Finally, in Section 7, we conclude.

2 Related literature

Our study relates mainly to two strands of literature: attitudes towards information and suboptimal decision-making in groups.

First, we contribute to the literature on attitudes towards information from both theoretical and empirical perspectives. On the theoretical side, standard economic theory recognizes that information is a valuable resource for decision-making and accordingly predicts that, if available at no cost, agents will acquire information about payoff-relevant states of the world to improve their subsequent decisions. However, a growing literature documents and acknowledges the possibility that economic agents may have intrinsic preferences for information beyond and above the instrumental, decision-value of information. Starting from [Kreps & Porteus \(1978\)](#), the theoretical literature has generated a variety of models in which agents may prefer to remain uninformed, or delay the acquisition of information, despite information being available at no or low cost. [Golman et al. \(2017\)](#) provides a review of models of information avoidance, defined as the deliberate choice to avoid freely available information when the agent is aware of its existence. [Bénabou \(2013\)](#), the model closest to our work, extends the analysis of information avoidance decisions to group settings. It shows theoretically that when willful blindness to information makes other agents worse off, it can become contagious. This raises the possibility of multiple equilibria, including a pervasive one in which all agents remain ignorant. We contribute to this literature by extending [Bénabou \(2013\)](#) to show that, under a broader class of preferences, strategic substitutability can also arise in that setting. In addition, we provide an empirical test of the prediction of interdependence.

On the empirical side, several studies both in the lab and in the field have documented that individuals often prefer to remain ignorant despite information being available at no or low cost

and being useful for subsequent decision-making (e.g., [Dana et al., 2007](#); [Karlsson et al., 2009](#); [Oster et al., 2013](#); [Ganguly & Tasoff, 2017](#); [Pagel, 2018](#); [Ho et al., 2021](#); [Meissner & Pfeiffer, 2022](#); [Masatlioglu et al., 2023](#); [Momsen & Ohndorf, 2023](#); [Engelmann et al., 2024](#); [Falk & Zimmermann, 2024](#)). Across studies and contexts, rates of avoidance range from 5% ([Ganguly & Tasoff, 2017](#)) to more than 90% ([Oster et al., 2013](#)). The vast majority of empirical studies have focused on information avoidance in individual decision-making settings. Most closely related to our paper, [Falk & Zimmermann \(2024\)](#) studies intrinsic preferences for information about a potentially adverse future event in an individual setting. In their experiment, information confers no instrumental value: the decision-maker’s future payoff is independent of whether information is acquired. We build on their experimental design to study information avoidance decisions in groups. Departing from them, we introduce a positive private return and externalities to information acquisition: becoming informed benefits the decision-maker and other members in her group. Therefore, we contribute to this literature in several ways. First, we document that, even in cases where information avoidance makes others worse off (negative payoff externalities), a substantial proportion of subjects still prefers to remain ignorant. Second, we provide, to the best of our knowledge, the first test of the interdependence of information decisions. Third, we document substantial heterogeneity of strategy types, and study its implications for the equilibrium levels of avoidance in groups. Fourth, we document how social preferences correlate with information decisions, which affect others’ wellbeing.

Second, our study also relates to the literature on non-Bayesian information processing, overconfidence, and suboptimal decision-making in groups. A recent literature suggests that decisions in groups can become correlated through the social transmission of biases. [Suvorov et al. \(2024\)](#) shows that members of a group share noisy feedback about their own ability to other group members in a selective and asymmetric way: group members are more likely to share positive feedback about their ability to perform in an investment task than negative feedback. This selective information sharing leads to groups becoming overconfident about the group’s ability to perform in investment decisions, and this overconfidence leads to suboptimal investment decisions by low ability groups. A closely related study by [Oprea & Yuksel \(2022\)](#) shows that when individuals can socially exchange ego-relevant beliefs, their initial biases amplify because subjects respond to their counterparts’ beliefs in an asymmetric way. Such studies have focused on ex-post processing of information, after its content has been revealed to the decision-makers. Complementing these studies, we focus on ex-ante avoidance of information, and show that groups can also attain suboptimal (material) outcomes by deliberately avoiding learning from freely available signals, when signals reveal the state perfectly, and without any communication among group members.

3 Theoretical framework

We now describe our theoretical framework and key results. All proofs are in the Appendix Section [A.1](#). Consider the following setup, inspired by ([Bénabou, 2013](#)). There are three periods, date 0, 1 and 2 and two states of the world, B (bad), occurring with probability $p \in (0, 1)$ and G (good), occurring with probability $1 - p$. Material payoffs are received at date 2 and

generate utility u_i . At date 1, agents evaluate lotteries over date-2 outcomes according to the expected utility function $U_i = E[u_i]$. At date 0, agents evaluate lotteries over date-1 utilities U_i according to the expected utility function $E[v_i(U)]$, where $v_i(\cdot)$ is a strictly increasing function capturing i 's preferences over expected utility lotteries in the spirit of Kreps-Porteus (1978). We may interpret $v_i(E[u_i])$ as an anticipatory utility experienced in date-1 over date-2 payoffs. For example, if $v_i(\cdot)$ is concave and the agent has interior priors $p_i \in (0, 1)$, then the loss in anticipatory utility upon learning that the state is Bad with certainty outweighs the gains in elation upon learning that the state is Good with certainty, creating an incentive to avoid learning the state. Denote with g the utility obtained at date 2 when the state of the world is good and with b when the state of the world is bad. b depends on whether an individual chooses to acquire or to avoid information about the state of the world at date 0. Intuitively, individuals who are aware that the state is bad can take corrective measures to reduce its negative consequences, whereas those who remain unaware cannot.² Thus, the utility at date 2 when the state is bad is equal to

$$b \equiv \bar{b} - \lambda_{-i} - \Delta \text{ if } i \text{ avoids information}$$

and

$$b + \Delta = \bar{b} - \lambda_{-i} > b \text{ if } i \text{ acquires information}$$

where $\bar{b} < g$ is the payoff when the state of the world is bad and all group members acquire information (and undertake corrective measures as a result), $\lambda_{-i} \in [0, \frac{n-1}{n}]$ is the share of group members other than i who avoid information, n is the number of individuals in the group and $\Delta \equiv \frac{1}{n}$. In words, Δ captures the instrumental value of acquiring information. It reflects how much an agent can lessen the adverse effects of a bad state by obtaining information and taking appropriate corrective measures. Note that, when the state of the world is bad, utility is decreasing in λ_{-i} , the share of i 's fellow group members who avoid information. By avoiding information, agents not only damage themselves but also exert a negative externality on others.

In what follows, we assume that

$$(1 - p)(g - \bar{b}) > \Delta. \tag{1}$$

This assumption is not necessary for our results but simplifies the exposition. In words, it ensures that the expected value of a lottery where i obtains $\bar{b}$ with probability p and g with probability $1 - p$ exceeds the value of obtaining $\bar{b} + \Delta$ for sure.³

An individual's date 0 *net* expected utility from avoiding information is

$$\varphi_i \equiv v_i(pb + (1 - p)g) - [pv_i(b + \Delta) + (1 - p)v_i(g)] \tag{2}$$

where $v_i(pb + (1 - p)g)$ is the date-0 expected anticipatory utility from avoiding and $pv_i(b +$

²This is imposed exogenously for simplicity but could arise endogenously in a setup where taking corrective action is costly, so that under ignorance agents would choose not to exert any corrective measure.

³Formally, $p\bar{b} + (1 - p)g > \bar{b} + \Delta$.

$\Delta) + (1 - p)v_i(g)$ is the expected anticipatory utility from acquiring information. If expression (2) is positive, individual i strictly prefers to avoid information at date 0, while if expression (2) is negative, individual i strictly prefers to acquire it.

3.1 Types of best response (“Strategy types”)

We now explore how the nature of v_i determines an individual’s optimal information acquisition choice as a function of others’ choices (i.e., her best reaction function). The first two results consider v_i that is everywhere convex or everywhere concave.

Lemma 1 *If v_i is everywhere convex, individual i is an “Always Getter”. At date 0, she acquires information independently of the information acquisition choices of other agents in her group.*

If v_i is convex, the utility boost obtained from discovering that the state is good outweighs the utility drop when the state is bad. Consequently, the agent always prefers to acquire information. Consider now the case where v_i is concave, as in Bénabou (2013).

Lemma 2 (Bénabou, 2013) *If v_i is everywhere concave, individual i (i) is an Always Getter (ii) is an “Always Avoider” or (iii) is affected by the choices of others in her group in the following way: there exists an interior threshold value λ_i^* such that i strictly prefers to avoid information if $\lambda_{-i} > \lambda_i^*$ and strictly prefers to acquire information if $\lambda_{-i} < \lambda_i^*$ (Strategic Complementarity).*

In the Appendix, we characterize the necessary and sufficient conditions for each of (i), (ii) or (iii) to apply. Always Getters and Always Avoiders make the same information acquisition choice independently of the choices of others in their group. Alternatively, the agent may condition her decision on the choices of others in her group, exhibiting strategic complementarity. The intuition for the result is as follows. Without the instrumental value of information, the agent would always prefer ignorance, since the utility loss from discovering the bad state outweighs the gain from finding out that the state is good (this follows directly from concavity). In the presence of instrumental value, as in our setup, the agent may opt to acquire information. However, as more agents choose to avoid information, the outcome in the bad state deteriorates, increasing the utility loss associated with discovering the bad state. As a result, meeting the conditions for information acquisition becomes increasingly challenging. This pattern may give rise to strategic complementarity in the agent’s information acquisition strategy. The agent acquires information only if a sufficient share of her peers also do so; otherwise, she avoids information.

Consider now a reference-dependent v_i (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). The agent’s utility is largely insensitive to expected payoffs except around some reference level. Denoting i ’s reference expected utility as $\widehat{U}_i$, we have (i) $v_i''(U) < 0$ for $U > \widehat{U}_i$ and (ii) $v_i''(U) > 0$ for $U < \widehat{U}_i$.

Lemma 3 *If v_i is reference dependent, under some conditions the following applies: there exists an interior threshold value λ_i^{**} such that i strictly prefers to avoid information if $\lambda_{-i} < \lambda_i^{**}$ and strictly prefers to acquire information if $\lambda_{-i} > \lambda_i^{**}$ (Strategic Substitutability).*

In the Appendix, we provide sufficient conditions for strategic substitutability. To illustrate, assume that if the state is known to be bad utility is always below the reference level while it is above it in the good state. First, suppose that many agents in i 's group avoid information. The final payoff in the bad state is so low that, at date 1, the expected utility under ignorance falls below the reference level. Relative to remaining ignorant, the utility loss from discovering the bad state is minimal (since both situations generate utility that is below the reference level), while the gain from discovering the good state is large. As a result, the agent prefers to acquire information. Now, suppose that many agents acquire information. The final payoff in the bad state is not too low and the expected utility under ignorance is above the reference level. In this case, under mild conditions, the agent prefers to avoid information, since the utility gain from discovering the good state is minimal relative to the utility loss from discovering the bad state.⁴

The following proposition summarizes our results.

Proposition 1 *Depending on the nature of v_i , i 's propensity to avoid information may be (i) independent of, (ii) increasing (strategic complementarity) or (iii) decreasing (strategic substitutability) in the share of group members who avoid information.*

In what follows, we will use the term “strategy type” to indicate whether an individual is an Always Getter, an Always Avoider, or alternatively exhibits preferences featuring strategic complementarity (in short, is a “Complement”) or substitutability (is a “Substitute”).

Naturally, the specific information avoidance equilibria that may be reached vary depending on the agents' strategy types. Bénabou (2013), for example, examines the risks that emerge when agents are Complements. In that case, groups may become trapped in a “Mutually Assured Ignorance” (MAI) equilibrium where no one seeks out information, and as a result, no corrective actions are taken to shield the group from adverse outcomes in the face of a bad state of the world —potentially with disastrous consequences.⁵ By the same token, a “Mutually Assured Awareness” (MAA) equilibrium, where everyone acquires information, is also possible. The path taken depends critically on the agents' expectations about the actions of others in their group. Conversely, a group of Substitutes cannot fall into either a MAI or MAA equilibrium, as its agents are inherently inclined to resist such dynamics.⁶

We examine the empirical existence of strategy types by designing an experiment that random varies beliefs about others' information choices, as well as by using the strategy method to elicit full best response schedules; results are discussed in Section 5. We also assess how the mixture of types impacts aggregate equilibrium avoidance by employing simulations in Section 6.

⁴For example, suppose $b(\lambda_{-i} = 0) = -10$, $b(1) = -90$, $g = 50$, $\Delta = 5$, $p = 0.5$, $\hat{u}_i = 0$, and $v_i(u) = u^{1/3}$ if $u > 0$ and $-(u)^{1/3}$ if $u < 0$. Then $\varphi^i(0) \approx 2.58$ and $\varphi^i(1) \approx -2.87$, so the agent exhibits strategic substitutability.

⁵Bénabou (2013) uses the acronym MAD (Mutually Assured Delusion) equilibrium. However, in the context of information avoidance, MAI provides a more accurate description of the equilibrium. The same holds for the acronym MAA (Mutually Assured Awareness).

⁶Consider, for instance, a group composed of Substitutes with a 0.5 threshold. It is easy to show that, in the unique equilibrium, exactly half the group acquire information and half avoid it.

4 Experimental design and procedures

4.1 Experimental design

We investigate whether, in groups, information avoidance is interdependent among people. Specifically, we ask: do individual decisions to remain uninformed depend on the expectation that others will do so?

4.1.1 Decision environment

To study the interdependence of avoidance, we design an experiment in which individuals in a group choose whether to acquire or avoid information. After being randomly assigned to groups, participants are told that, at the end of the experimental session, a state of the world (either good or bad) will be randomly drawn. In the bad state (or the “Screams” state), the group will experience a negative consumption event: listening to a series of unpleasant screams. In the good state (or the “Quiet” state), no such event happens.

Each member of the group can choose one of two options, at no monetary cost. If she chooses option “Now”, she learns the state of the world immediately, while if she chooses option “Later” she defers learning the state of the world until the consumption event takes place at the end of the experimental session.⁷ The choice between “Now” and “Later” is the key decision of interest. We refer to this binary choice as the “information decision”.⁸

A feature that distinguishes our experiment from other studies is that information decisions generate a payoff externality: the acquisition of information creates a positive return not only for oneself (standard instrumental value of information), but also for other members of the group (externality). If the state is *Screams*, the volume of the screams that all group members will hear decreases in the total number of group members who acquire information (option Now). More precisely, each individual knows that by choosing Now, she lowers the volume of the screams in the Screams state by 2 points *for all* group members.⁹

In our decision environment, the acquisition of information is costless under standard accounts: there are no monetary costs associated with choosing either Now or Later. In addition, cognitive costs of processing information are negligible, as upon acquisition subjects learn only a binary state of the world.

4.1.2 Treatments

We aim to identify whether information decisions are interdependent. To this end, we designed our experimental treatment to induce exogenous variation in participants’ beliefs about the

⁷Participants are aware that if they choose Later, they will be informed about the outcome just before the period in which screams are potentially played.

⁸In each group, decisions are made simultaneously and without communication. As a result, there is no scope for social learning among group members.

⁹As described in Section 4.2, the contribution to the volume of screams is additively separable. Therefore, we examine whether strategic interdependence in information decisions can arise even in the absence of any in-built interdependencies in the technology.

share of information avoiders (i.e., people who would choose Later) in their group. We do so through an information provision treatment (Haaland et al., 2023), where we randomize information about the share of information avoiders in the past. There are two treatment conditions, labeled the *Many Avoided* and *Few Avoided*, randomized at the group level. In *Many Avoided* (*Few Avoided*), we inform that 80% (20%) of subjects in a particular group of a past pilot similar to the current experiment chose to avoid information about the state. Participants are also reminded of the potential impact on scream volume if their group members exhibit similar information avoidance behaviors as those observed in the pilot study.¹⁰ To check whether the treatment manipulation works as intended, after the information provision treatment we elicit the participants’ beliefs about the fraction of avoiders among their group mates.

To assess the (average) interdependence of information decisions, we compare the fraction of information avoiders (i.e., people who chose option Later) between *Many Avoided* and *Few Avoided*.

4.1.3 Individual best-response schedules and heterogeneity

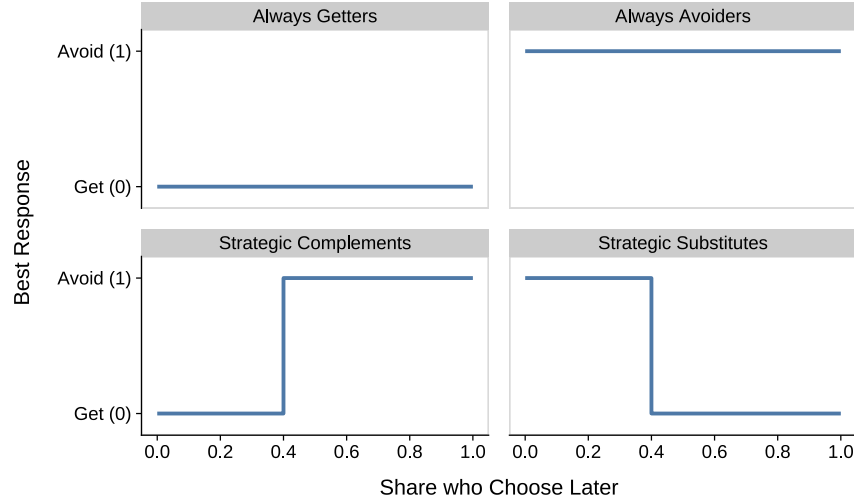
Proposition 3.1 suggests that individuals can be heterogeneous in how they respond to others’ information decisions. To delve into this heterogeneity, we use the strategy method (Selten, 1967) to examine how each participant responds to a range of choices of their group mates, in an incentive compatible way. We elicit the full schedule of each individual’s best response to all the possible choices made by others in the group. We classify the best reaction schedules based on whether a participant chooses Later when the share of avoiders among other members is sufficiently large (Strategic Complement) or sufficiently small (Strategic Substitute). If the participant’s choice is the same for all distributions of information avoidance in their group, we classify them as Always Getter (if they always choose to acquire information) or Always Avoider (if they always choose to avoid information). The remaining cases (subjects who switch multiple times) are classified in a residual category labeled “Others”. We refer to these categories of best responses as “strategy types”, paralleling the theoretical classification defined in Section 3.1. Figure 1.1 illustrates examples of best reaction functions for each strategy type.

4.1.4 Additional measurements

Baseline attitudes towards information To measure baseline preferences over information, we administer the “Information Preferences Scale” (“IPS”; Ho et al., 2021). The IPS captures an individual’s willingness to obtain information that may be unpleasant but is instrumentally valuable, across multiple domains. Higher IPS scores indicate greater general tendency to seek information. Measuring these general preferences serves three purposes: (i) to verify balance between treatment groups on a trait potentially relevant for information decisions; (ii)

¹⁰Specifically, we inform each participant that, if their group mates behave like the participants of the pilot, the volume in the *Screams* state would be, depending on whether she herself avoids or acquires information, 89 or 91 (in the *Many Avoided* condition), and 60 or 62 (in the *Few Avoided* condition). The difference of 2 points within each condition arises from the focal participant’s choice. Conditional on her choice, the difference in scream volume between conditions arises from the information choices of others.

Figure 1.1: Examples of best-response functions, by strategy type



Note: This figure illustrates examples of best-response functions, categorized by “strategy type”. These types, developed from the model, classify best response schedules by whether they are independent from others’ choices (“Always Getters” and “Always Avoiders”) or conditional (“Strategic Complements” and “Strategic Substitutes”). The x-axis represents the share of other group members avoiding information. The y-axis indicates the individual’s choice to acquire or avoid information. Blue lines depict best reaction schedules. Vertical lines indicate example switching thresholds for Strategic Complements and Strategic Substitutes (actual thresholds may vary across individuals).

to include as a control, reducing residual variance and improving the precision of our estimates; and (iii) to benchmark the predictive power of our elicited anticipatory utility measures (see below).

Mediators: Anticipatory utility To shed light on mechanisms and determinants of information decisions, we introduce a novel elicitation designed to measure anticipatory utility over the future consumption event (screams or silence). A crucial feature of our elicitation is its conditionality: anticipatory utilities are measured *conditional on each information set*—namely, upon acquiring information and discovering a positive signal (“good news” event), upon acquiring information and discovering a negative signal (“bad news”), and avoiding information and discovering no signal at all (“no news”). Specifically, we ask participants:

- a) “How happy would you be *thinking about the outcome ahead* if you chose Now and then discovered that the lottery outcome is Screams?”;
- b) “How happy would you be *thinking about the outcome ahead* if you chose Now and then discovered that the lottery outcome is Quiet?”; and
- c) “How happy would you be *thinking about the outcome ahead* if you chose Later?”

The elicitations above attempt to measure, respectively, the anticipatory utilities $v_i(b + \Delta)$, $v_i(g)$, and $v_i(pb + (1 - p)g)$. To compute the net utility of avoiding information φ_i , we also elicit p_i , the participant’s subjective belief that the state is Screams (i.e., the prior belief of

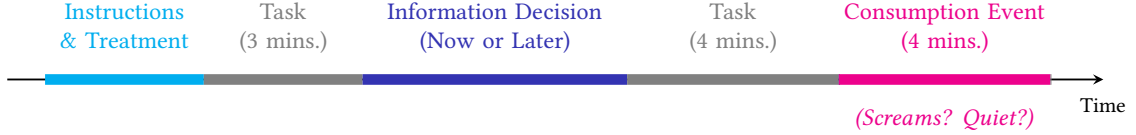


Figure 1.2: Timeline of experiment

finding out bad news). Together, these four questions mirror the components of expression (2), which defines the net utility of avoiding versus acquiring information and theoretically determines the information decision. This granular elicitation of the components of φ_i allows us to examine how the treatment shifts the expected anticipatory utility of learning good versus bad news separately, which helps to distinguish among mechanisms. We call this set of questions the “Kreps-Porteus questions” (or “KP questions”).

Social preferences Since information decisions generate payoff externalities on other members, prosociality may be a relevant predictor of preferences for information. To empirically investigate the relationship between social preferences and information decisions, we elicit individual measures by administering the Global Preferences Survey (GPS) (Falk et al. (2018)), which captures altruism and reciprocity, as well as trust, risk attitudes and time preferences.

4.1.5 Task and earnings

Following the design in Falk & Zimmermann (2024), we include an incentivized task for participants to complete in the first part of the experimental session, after the participants have taken the Now or Later decision and before the second part of the experiment where they may or may not experience the screams depending on the realization of the state. The experimental timeline is summarized in Figure 1.2.

4.2 Experimental procedures

The experiment was conducted at the Centre for Decision Research and Experimental Economics (CeDEx) Lab, University of Nottingham.¹¹ We ran 17 experimental sessions (each constituting an independent group) across two waves of data collection (December 2023 and May–June 2024), recruiting participants from the CeDEx subject pool. In total, the sample for analysis comprised 465 participants, with an average of 27.35 (± 3.66) subjects per session.¹² Each session lasted about one hour, and participants earned on average £12.52 (± 2.00), including a £5 show-up fee. The average score on the Information Preferences Scale (IPS) was 3.12 (± 0.41),

¹¹The experiment was preregistered, AsPredicted #153871 and #175588. The experiment was approved by the University of Nottingham’s School of Economics Research Ethics Committee. At the start of the session, participants were informed about the nature of the experiment and provided written consent to participate.

¹²Originally, we recruited 493 participants in 18 sessions. However, due to technical issues, data (including the main outcomes of interest) was not recorded for 22 subjects in a single session of 28 participants and is excluded from the analysis; our results are robust to the inclusion of the 6 individuals from this session for which data is available. In a separate incident, one participant left early, about an hour after the session started, reducing observations for final-stage variables by one.

indicating a general tendency to seek potentially undesirable information.¹³ Full descriptive statistics of baseline characteristics are reported in Table A.1 in the Appendix.

At the start of each experimental session, participants were matched into groups, and each group was randomly assigned to a treatment condition.¹⁴ Participants were then provided with detailed instructions on computer screens.¹⁵ The instructions included a description of the state of the world lottery, the negative consumption event (screams), the timeline of the experiment, the information decisions and their implications, and the task. In order to ensure understanding of the decision environment, we included control questions about the timing of learning the outcome and about the nature of returns and externalities from information acquisition/avoidance. During the instructions phase, we also implemented our information provision treatment.

A computerized lottery determined the outcome (either *Screams* or *Quiet*) for each group. If the outcome was *Screams*, in the second part of the experiment the members of the group listened through headphones (worn throughout the session) to a series of unpleasant screams, drawn from the validated stimulus set in Beurenaut et al. (2020).¹⁶ To ensure hearing safety, the maximum possible volume level (100) was calibrated at 80dB in accordance with UK hearing safety regulations. The calibration was carried out separately for each headphone-computer pair, ensuring the hearing safety of every subject. To ensure exposure to the potential screams, participants were informed at the outset that headphone use was mandatory throughout the session and that attempts to remove them would be a basis for expulsion.¹⁷ Compliance and correct headphone use were monitored by two invigilators per session. Participants were told they could leave the session at any time.¹⁸

The use of screams instead of electric shocks (employed in experimental studies on information avoidance in individual decision making) is based on three rationales. First, one of the motives that may lead to information avoidance are anticipatory emotions, such as anxiety about a future outcome, and studies show that the screams are able to sustain anxiety for longer periods than electric shocks (Beurenaut et al. (2020)). This feature is especially desirable for experiments with group decision-making, which typically take longer compared to individual decision-making. Second, the provision of human scream sounds is more scalable than electric shocks, since the only required equipment are computers and headphones, as opposed to devices specialized for the provision of shocks.¹⁹ Non-scalability imposes a natural limit to group sizes, which increase the likelihood of the emergence of complementarities in

¹³A score of 3 means “Probably want to know”.

¹⁴The random assignment of the treatment was blocked by week of year and time of session (morning and afternoon).

¹⁵Full instructions are available in Appendix Section A.3. The experiment was programmed in Lioness (Giamattei et al., 2020).

¹⁶Participants did not hear any sample screams before making their information decisions, but were informed that “[p]revious research has shown that people consider similar screams as *more aversive (disturbing) than the sound of nails sliding on a chalkboard*”. Evidence of this statement can be found in Kumar et al. (2009).

¹⁷We are grateful to the people who generously helped as invigilators: Pierce Gately, Kieran Stockley, Tong Fang, Xue Wang, Matías Golman, Jesús Rodríguez, Yifan Li, Adrian Brown, Thomas Barnes, Lara Suraci, and Niramay Chugh.

¹⁸Only one participant left, after about an hour of the session, reportedly due to the start of a lecture.

¹⁹The lower implementation barrier also facilitates replicability.

information decisions (Bénabou (2013)). Third, screams allow a more granular manipulation and more intuitive understanding of variations in severity (i.e., volume levels as opposed to electric shock intensity); this granularity in severity is important as group sizes become larger and the marginal return of acquisition of information (Δ) by any individual member becomes smaller. Fourth, the use of human screams alleviates ethical concerns compared to electric shocks.²⁰

The experimental and model assumption is that screams are a bad. To assess whether this is the case, we elicit the participants' utility over two different levels of volume of screams. We ask: "Imagine that the Outcome is Screams, and that the volume of the screams will be level 100. How happy would you be thinking about the Outcome ahead?"; we ask a similar question for volume level 50. This allows us to identify subjects who, contrary to our assumption, derive higher utility from higher levels of volume of scream sounds ("volume lovers").

For groups whose lottery outcome was *Screams*, the volume of the screams was determined by the information decisions of its members. Specifically, the volume of the screams in group g was determined by the following function:

$$\text{vol}_g = 50 + (100 - 50) \left(\frac{1}{N_g} \sum_{i=1}^{N_g} a_{ig} \right), \quad (3)$$

where $\text{vol}_g \in [50, 100]$ is the volume of screams for group g conditional on the lottery outcome being *Screams*, N_g is the number of members in group g , and $a_{ig} = 1$ if subject i in group g chooses *Later* and 0 otherwise. The acquisition of information about the state, $a_{ig} = 0$, thus partially mitigates the volume of the screams.

While the volume production function in Equation (3) appears related to a public goods game (PGG), our decision environment differs in several ways that reverse the standard theoretical predictions. First, the nature of the cost is distinct: in a PGG, the cost of contributing is a known monetary amount, identical for all players and independent of others' actions. Here, there is no monetary cost of contributing: the cost is instead psychological (from potentially learning bad news about the future) and is heterogeneous across individuals. Moreover, the psychological cost plausibly varies with the choices of others (which influence how severe the bad state is). These differences lead to a stark divergence in the prediction for a purely material-payoff maximizing agent: such an agent would optimally free-ride in a standard PGG but would always prefer to "contribute" (i.e., acquire information) in our setting, a dominant strategy for maximizing material outcomes. Consequently, the central puzzle in each setting

²⁰Recently, experimental literature in Economics has started using the prospect of money losses as a negative future event (Pagel, 2018; Engelmann et al., 2024). In particular, Engelmann et al. (2024) finds that money losses induce anticipatory anxiety and wishful thinking about the future event. The literature on intrinsic preferences for information has focused attention on future *consumption* events as distinct from *income* events, because knowing the latter introduces a planning advantage that motivates the acquisition of information about the future income event. To avoid introducing this extra planning advantage, in our experiment, we follow the latter tradition and focus on a consumption event. We argue that, theoretically, the relevant feature of the event is that the prospect of hearing human screams, much like the prospect of money losses, is perceived as a bad event by participants. In Section 5.1.1, we show evidence that, in our experimental sample, this is the case.

is inverted: whereas the literature on PGGs seeks to explain why people *do* contribute, in our setting the question is why people *do not* (i.e., why they avoid information) and how this reluctance to contribute is shaped by others' decisions.

Each group was composed of an entire experimental session, with group sizes of approximately $N_g = 30$. This implies that the return to each participant's information acquisition (in terms of volume reduction) was $1/30 * 50 = 1.67$ volume points.²¹

The information decision (the binary decision between options Now and Later) is our main outcome of interest. To obtain an alternative, continuous measure of intensity of preferences, we also ask participants, after they chose either Now or Later, to rate the strength of preference for their selected option on a Likert scale ranging from 0 ("I am indifferent") to 10 ("I have a very strong preference for the selected option"). From the measure, we construct a continuous version of the information decision, weighted by the reported strength of preference. This continuous outcome variable ranges from -10 ("I have a very strong preference for option Now") to 10 ("I have a very strong preference for option Later"). We also ask participants to write down the reason for the selected option in an open text box.

To explore individual heterogeneity in best responses, we use the strategy method (Selten, 1967) to elicit each participant's complete schedule of choices between options Now and Later as a function of others' choices. This schedule captures their choice conditional on every possible share of information avoiders among others in the group, specifically for the shares: 0%, 20%, 40%, 60%, 80%, and 100%.

We make the strategy method elicitation incentive-compatible by assigning a strictly positive probability that these choices are implemented. Specifically, before the elicitation, we inform participants about the following procedure. After the group members make their (unconditional) informational decisions (between Now and Later), we randomly select $N_g - 1$ subjects and calculate the proportion of information avoiders *among these* $N_g - 1$ subjects, based on their unconditional decisions. For the remaining participant, we implement her selected best response to this share.

To measure general preferences for information, we administer the "Information Preferences Scale", or "IPS" (Ho et al., 2021). This scale measures an individual's desire to obtain or avoid information that may be unpleasant but is instrumentally valuable, across several domains.²² The scale ranges from 1 = "Definitely not want to know" to 4 = "Definitely want to know". Therefore, higher IPS scores indicate greater general willingness to obtain information.

To measure social and economic preferences, we administer the Global Preferences Survey (GPS) (Falk et al., 2018). The GPS measures risk attitudes, time preferences, altruism, reciprocity, and trust. Each dimension of preferences is elicited through the responses to two questions.

²¹We described the return to information acquisition to participants as being equal to "around 2 out of 100 volume level points".

²²The scale is constructed by averaging the answers to 18 questions that ask the subject's willingness to obtain information in hypothetical scenarios. The scenarios involve five domains: health, finance, social relations, ego-relevant characteristics, and occupation. For each question, the 4 possible answers are "Definitely don't want to know" (encoded as 1), "Probably don't want to know" (2), "Probably want to know" (3), and "Definitely want to know" (4).

Following the procedure in [Falk et al. \(2018\)](#), we standardize the answers, transforming them into z-scores.²³ Higher z-scores indicate, respectively, higher risk seeking, patience, altruism, reciprocity, and trust. We also elicited demographics (e.g., age, gender), prior work experience, ratings about the perceived difficulty of instructions and quiz questions. Participants were given the opportunity to feedback on their experience in the experiment in open text boxes.

Following [Falk & Zimmermann \(2024\)](#), we include an incentivized task after the Now or Later decision and before the second part of the experiment. Participants are asked 90 general knowledge quiz questions, with earnings increasing in the number of correct answers. Each participant is paid according to her own performance, which is the only source of monetary earnings in our experiment, apart from the show-up fee.

5 Results

This section presents the analysis of our experimental data. We begin by establishing two necessary preconditions for our study. In Section 5.1.1, we verify that, for most participants, screams are perceived as a bad, confirming our theoretical and experimental assumption. In Section 5.1.2, we document the existence of information avoidance in our decision environment (a context where the positive instrumental value and externalities from acquisition of information create an incentive to *acquire* it). Further, we establish that the decision to avoid information in our environment is likely not random nor driven by indifference, as avoiders report strong preferences for their choice.

Following these prerequisite validations, we turn to our primary results. In Section 5.2, we leverage experimental variation in beliefs about others' information decisions to test for strategic interdependence, assessing whether the likelihood of information avoidance increases (strategic complementarity) or decreases (strategic substitutability) on average. In Section 5.3, we move beyond the average to characterize individual heterogeneity. Using the strategy method, we document the empirical distribution of strategy types and report heterogeneous treatment effects by type. In Section 5.4, we explore the mechanisms driving the average treatment effects. Specifically, we use the KP questions to decompose the effect by identifying which components of the expression for the net incentive to avoid information (2) are most impacted by the treatment. We also discuss the plausibility of alternative mechanisms. Finally, the next section is devoted to analyzing how the group composition, in terms of strategy types, shapes the distribution of equilibrium avoidance.

²³For risk attitudes and time preference, one of the questions used by [Falk et al. \(2018\)](#) involve the staircase method. Since administering this could be time-consuming and cognitively burdening in the middle of our experiment, we did not include such questions.

5.1 Preliminary analyses

5.1.1 Are screams a bad?

Before presenting our primary findings, we first conduct a validation check of a foundational assumption underlying our study: that participants actually perceive the prospect of listening to screams as a bad. We provide evidence confirming that the assumption holds for our sample. First, we asked subjects how “(un)happy” they would be if screams were played at volume 50 and, separately, at volume 100. The difference in these ratings provides a proxy of each participant’s disutility of listening to screams at higher volume. Comparing these ratings reveals that a large majority of subjects report lower utility levels for screams at higher volumes: specifically, 82% of subjects report that screams at volume 100 would be strictly worse than at volume 50.²⁴ Second, using the KP questions, we asked subjects to rate their “(un)happiness” upon discovering that the state was Screams after choosing Now, and separately, if it was Quiet. The difference in these ratings provides a measure of each participant’s utility for listening to screams relative to silence. If the screams are perceived as a bad, we would expect lower ratings for the Screams state than for the Quiet state. As expected, this pattern holds for the majority of subjects: 81% report that discovering that the Screams state is strictly worse than discovering that the Quiet state.²⁵ The distribution of types of preference screams and volume is shown in Figure A.3 and Figure A.1 in the Appendix. The evidence hence supports the view that the vast majority of subjects perceive screams as a bad.

Furthermore, the fact that higher-volume screams are perceived as worse than lower-volume ones is compatible with the notion that participants perceive the instrumental value of information acquisition (which allows them to reduce the volume of screams) as non-trivial.

For our main analysis, we consider all subjects, including those with weak preferences for screams and volume (or both). We refer to such subjects respectively as “scream-lovers” and “volume-lovers”.²⁶ Analyses that exclude these individuals are presented in the Appendix Section A.2.4. They confirm that our findings are robust to the exclusion of volume- and scream-lovers, and, in some cases, become even stronger.

5.1.2 How prevalent is information avoidance in groups?

In order to study the interdependence of information avoidance, it must first be established that avoidance occurs. We assess this by focusing on the information decision —the binary choice between learning the state Now or Later. The decision is coded as $a_{ig} = 1$ if individual i in group g chose Later and 0 otherwise. In addition, we construct a continuous measure, the “information avoidance scale” (“IA scale”), which captures the stated *strength of preference* for

²⁴Among the remaining subjects, 3% are indifferent between volume levels, and 9% would prefer volume 100. We call the latter type of subjects “volume-lovers”.

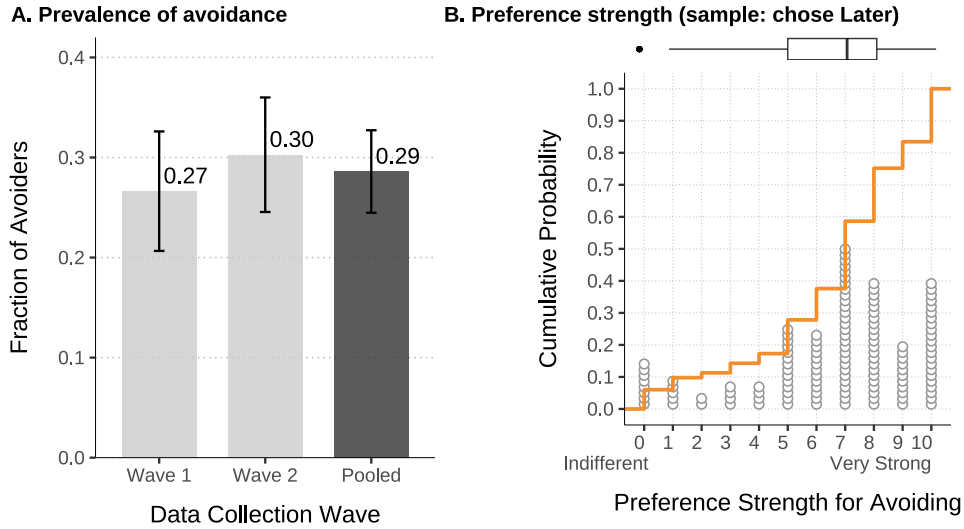
²⁵Among the remaining subjects, 6% are indifferent between the two states and 14% would prefer discovering the Screams state. We call the later type of subjects “scream-lovers”.

²⁶There are 111 subjects who are either scream lovers or volume lovers or both. The distribution of preferences is reported in Appendix Section A.2.2.

either option.²⁷ Analyses of both measures yield qualitatively similar results. For parsimony, we report findings for the binary decision in the main text and defer the IA scale results to Appendix Section A.2.6.

We find statistically significant evidence of information avoidance. Panel A of Figure 1.3 shows its incidence by data collection wave and for the pooled sample. In the pooled data, the incidence is 0.286, significantly greater than zero at 1% level. The proportion was stable across waves.

Figure 1.3: Information avoidance: incidence and preference strength



Note: The figure characterizes the incidence of information avoidance in our experimental setup. **Panel A** shows the fraction of subjects who chose Later (avoided information). The first two bars (brown) show the fraction by data collection wave; the third bar (blue) shows the fraction in the full sample. Whiskers indicate 95% confidence intervals. Pairwise comparisons using two-sample proportion tests yield p-values greater than 0.4. **Panel B** restricts attention to the subset of subjects who chose Later (avoided information), showing the distribution of the strength of preference for this choice. Preference strength was elicited on a Likert scale from 0 (“I am indifferent” between Now and Later) to 10 (“I have a very strong preference” for the chosen option). Gray dots represent the absolute frequency of each preference strength (one dot per subject). The orange line shows the empirical cumulative distribution function (CDF). The boxplot shows the median and the interquartile range, with whiskers extending to the most distant point within 1.5 times the interquartile range.

This result relates to prior findings in the literature. In an individual decision-making context, [Falk & Zimmermann \(2024\)](#) find that 52% of subjects avoid information. The higher proportion in their study could be attributed to a series of factors. In addition to some differences in implementation, in their setup information provides no material benefit (i.e., it does not affect the severity of the future negative consumption event), eliminating instrumental and prosocial motives for information acquisition.

In our data, avoidance does not appear to be driven by indifference or trembling. Panel B

²⁷After making their information decisions, participants rated the strength of their preference for the selected option from 0 (“I am indifferent”) to 10 (“I have a very strong preference”). For those choosing Later, the IA scale equals the reported strength; for those choosing Now, it equals the negative of that strength. Thus, the IA scale ranges from -10 (very strong preference for Now) to +10 (very strong preference for Later), with 0 indicating indifference, thereby providing more variation than the binary measure.

of Figure 1.3 plots the distribution of preference strength among avoiders (people who chose Later), elicited on a 0-10 scale (0=“indifferent”; 10=“very strong preference”). The average strength of preference is 6.6, well above indifference. Moreover, the cumulative distribution (orange line) shows that only 6% of avoiders report indifference, while more than 80% report a strength of at least 5. Thus, most avoiders strongly prefer to remain uninformed, rather than choosing randomly or out of indifference.

5.2 The interdependence of information avoidance: experimental evidence

Having established the two prerequisites, we now turn to our main question: whether information avoidance in groups is interdependent, which we examine leveraging the information provision treatments.

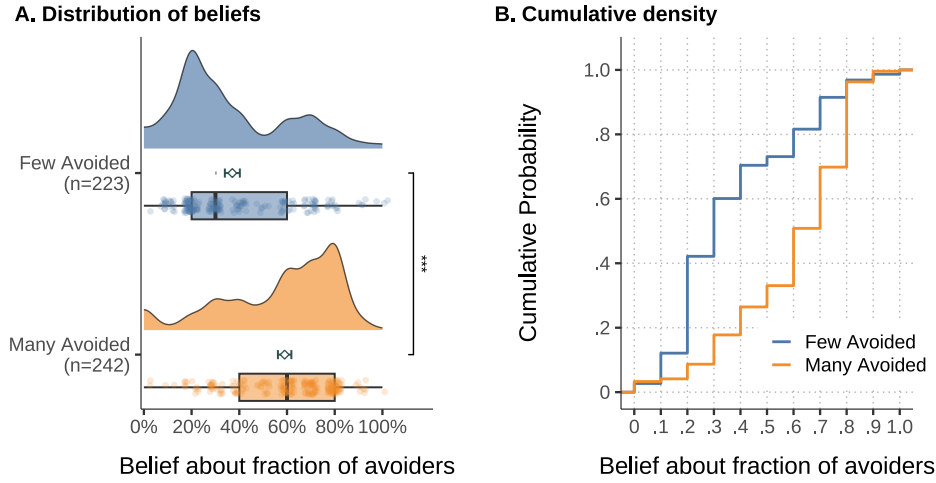
We first report treatment balance checks in Appendix Table A.3, showing that all baseline characteristics are balanced between conditions, except for the representation of some fields of study.²⁸ To account for this, we report results with and without controls for field of study in the regressions. Importantly, the IPS scores are balanced across treatment groups, for both the overall scale and for each of the domain-specific subscales. Therefore, any observed differences in information avoidance between treatments cannot be attributed to baseline imbalances in general or domain-specific attitudes towards information.

Our treatment manipulation successfully shifts expectations about the share of avoiders among other members in the group. Panel A in Figure 1.4 shows the average belief by condition. As intended, the information provision treatment substantially shifts beliefs upwards. The difference in average beliefs about peers’ informational decisions is large and significant: on average, subjects in *Many Avoided* believe that a 0.219 higher proportion of their peers would avoid information, compared to those in the *Few Avoided* condition. The difference represents a 59.12% increase relative to the *Few* condition, and is statistically significant at 1% level. Moreover, the shift in beliefs is not only in the mean: Panel B plots the empirical CDFs of beliefs by condition, showing that the distribution of beliefs in *Many Avoided* first-order stochastically dominates (with a small number of exceptions at the extremes of the distribution) the distribution in the *Few Avoided* condition. In sum, the experimental treatment was successful in shifting beliefs about others’ behavior, as intended.

Table 1.1 reports estimates from linear regressions of beliefs on the treatment dummy and different sets of controls. Column (1) reproduces the results in Figure 1.4: on average, the expected share of avoiders among other group members is 21.9 percentage points higher in *Many Avoided* than in *Few Avoided*. This difference is highly significant and robust to the inclusion of the following controls: strata fixed effects (column 2), field of study fixed effects and individual characteristics (age and gender; column 3), and economic and social preferences (altruism, reciprocity, risk-seeking, patience, trust), as well as the IPS (column 4). Overall, the evidence confirms that the treatment successfully shifted beliefs about the prevalence of avoidance among other group members.

²⁸Compared to *Few Avoided*, in *Many Avoided* there is a lower share of Business and of Economics students.

Figure 1.4: Manipulation check



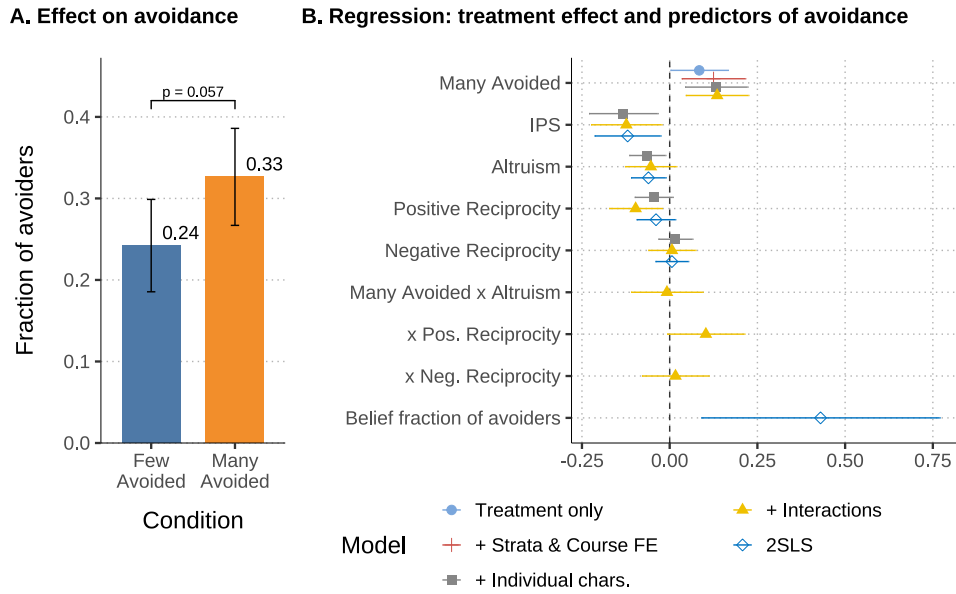
Note: The figure shows subjects' beliefs about the fraction of information avoiders among other subjects in their own group, by treatment condition. **Panel A** shows the average belief by condition. Error bars show 95% confidence intervals; p-value from a two-sample t-test. **Panel B** shows the empirical CDF of beliefs by treatment condition.

Having established that the treatment shifted beliefs about others' avoidance, we now examine whether it also generated differences in actual behavior. Panel A of Figure 1.5 plots the proportion of information avoiders by treatment condition. Compared to *Few Avoided*, the share of avoiders in *Many Avoided* is 0.084 higher ($p = 0.045$), representing a 34.81% increase. Regression estimates from a linear probability model are visualized in panel B of Figure 1.5, and reported in Table 1.2 for more detail. In the unconditional specification (blue circles, or column 1), the treatment effect is 0.084, significant at the 5% level. This effect is robust to the inclusion of individual characteristics and strata and field-of-study fixed effects (red crosses, column 2; $\hat{\beta} = 0.109$, significant at 5%), and increases to 0.123 when all controls are included, significant at the 1% level (gray squares, column 4).

The estimate of the treatment effect is robust to a range of alternative specifications. Excluding Scream- and Volume-Lovers from the sample makes the effects even stronger (Appendix Tables A.4 and A.5). We obtain qualitatively similar results when using the IA scale (the continuous preference strength for information avoidance) as the outcome (Appendix Section A.2.6). Finally, Appendix Section A.2.5 shows that the estimate remains robust and significant across alternative specifications, regardless of whether the set of controls is selected by the researcher or via the data-driven post-double selection procedure (Belloni et al., 2014).

To quantify the relationship between beliefs and behavior (the interdependence), we move beyond the reduced-form results and estimate a 2SLS model instrumenting beliefs with the

Figure 1.5: Information avoidance: treatment effect and predictors



Note: The figure shows the average treatment effect on information avoidance. In the *Few Avoided* condition, subjects were told that 20% of participants in the past avoided information (suggesting that volume would be around 60); in the *Many Avoided* condition, the analogous fraction was 80% (volume around 90). **Panel A** displays the fraction of avoiders by treatment condition. Error bars indicate 95% confidence intervals; p-value derived from a two-sample proportion test. **Panel B** shows coefficients from linear probability models where the dependent variable is a dummy for avoiding information (Later = 1, Now = 0). The first model (blue dots) includes only the treatment dummy. The second model (red plus signs) adds strata and field of study fixed effects. The third model (gray squares) further controls for individual characteristics: the Information Preferences Scale (Ho et al., 2021; with higher values indicating a stronger overall preference for *knowing* potentially undesirable information), gender, age, social preferences (altruism, positive and negative reciprocity), and economic preferences (risk aversion, patience, and trust; omitted from the figure for conciseness). The fourth model (yellow triangles) incorporates interactions between treatment and social preferences. The last model (blue hollow diamonds) instruments beliefs with the treatment assignment and shows 2SLS estimates. Error bars represent 95% confidence intervals computed based on robust standard errors.

Table 1.1: Manipulation check (regressions)

	Dep. var: Belief about share of avoiders			
	(1)	(2)	(3)	(4)
Many Avoided	0.219*** (0.021)	0.215*** (0.025)	0.218*** (0.025)	0.221*** (0.025)
Control group mean	0.371*** (0.016)	0.471*** (0.072)	0.394*** (0.137)	0.494*** (0.163)
Strata FEs		✓	✓	✓
Field of study FEs			✓	✓
Individual characteristics			✓	✓
Preferences (economic, social, IPS)				✓
Observations	465	465	460	460

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table shows the estimated coefficients of a regression of beliefs about the share of information avoiders among other members on the treatment dummy. The dependent variable is a continuous measure of beliefs, ranging from 0 (0% of the subject’s group will avoid information”) to 1 (100%). The treatment dummy is equal to one if the subject was in the *Many Avoided* condition and zero if in the *Few Avoided* condition. Individual characteristics include age and gender. “Preferences” include economic and social preferences (risk-seeking, time preferences, altruism, reciprocity, and trust) and the Information Preferences Scale (Ho et al., 2021), measuring an individual’s general preference for obtaining information. Field of study fixed effects include dummies for fields of study. Strata-fixed effects include dummies for week of year and time of day. The number of observations drops in columns (3) and (4) because five participants did not provide information about their age.

treatment assignment (Appendix Table A.8).²⁹ The estimated coefficient ranges from $\hat{\beta}^{2SLS} \approx 0.4$ to 0.6, implying that a 10 percentage point increase in expected avoidance by others raises an individual’s probability of avoiding information by about 5 percentage points. This supports the conclusion that, on average, information avoidance is a strategic complement in our sample.

The regressions in Table 1.2 also show that social preferences are correlated with information decisions and may moderate the interdependence. Column (4) includes social preferences, and column (5) their interactions with the treatment dummy. While the estimated coefficients cannot be interpreted causally, we offer descriptive evidence. For reciprocity, we observe a significant and positive interaction with the treatment: more reciprocal individuals are on average less likely to avoid in the *Few Avoided*, and more likely in the *Many Avoided* condition. Intuitively, avoiding information is a way of taking revenge by harming others, so more reciprocal individuals are less inclined to avoid information when only a minority avoids information

²⁹The exogeneity assumption is satisfied by design. The monotonicity assumption —requiring that for every subject potential beliefs are weakly higher in *Many Avoided* relative to *Few Avoided*— is likely to be satisfied, given the first-order stochastic dominance of the belief distribution in *Many Avoided* relative to *Few Avoided* shown in Figure 1.4. The exclusion restriction requires that the effect of the information provision treatment operates only through posterior beliefs about others’ information decisions. While this assumption is not directly testable, we discuss mechanisms in Section 5.4.

(*Few Avoided*) than when the majority does (*Many Avoided*). However, as Appendix Table A.4 and Table A.5 show, the estimated coefficients shrink when Scream- and Volume-Lovers are excluded, becoming statistically insignificant, although the signs of the point estimates remain unchanged. For altruism, our data indicate that more altruistic individuals are less likely to avoid information ($\hat{\beta} = -0.068$), consistent with the idea that prosocial motives encourage information acquisition to benefit others.

Lastly, the Information Preferences Scale significantly predicts behavior in our experiment: individuals with a stronger tendency to seek potentially undesirable information (higher IPS scores) are less likely to avoid information in our experiment.

In summary, our information provision experiment provides causal evidence that, in groups, information avoidance can be interdependent, and in particular, contagious.

Table 1.2: Treatment effect on information avoidance

	Dependent variable: Avoided information				
	(1)	(2)	(3)	(4)	(5)
Many Avoided	0.084** (0.042)	0.109** (0.049)	0.114** (0.049)	0.123** (0.049)	0.124** (0.049)
Information Preferences Scale			-0.125** (0.051)	-0.130** (0.052)	-0.118** (0.054)
<i>Social preferences</i>					
Altruism				-0.065** (0.028)	-0.052 (0.041)
× Many Avoided					-0.006 (0.057)
Positive reciprocity				-0.046 (0.029)	-0.101** (0.041)
× Many Avoided					0.109* (0.060)
Negative reciprocity				0.013 (0.026)	-0.005 (0.036)
× Many Avoided					0.030 (0.051)
Constant	0.242*** (0.029)	0.126 (0.225)	0.326 (0.301)	0.280 (0.311)	0.273 (0.321)
Strata FEs		✓	✓	✓	✓
Field of study FEs		✓	✓	✓	✓
Individual characteristics			✓	✓	✓
Observations	465	465	460	460	460

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

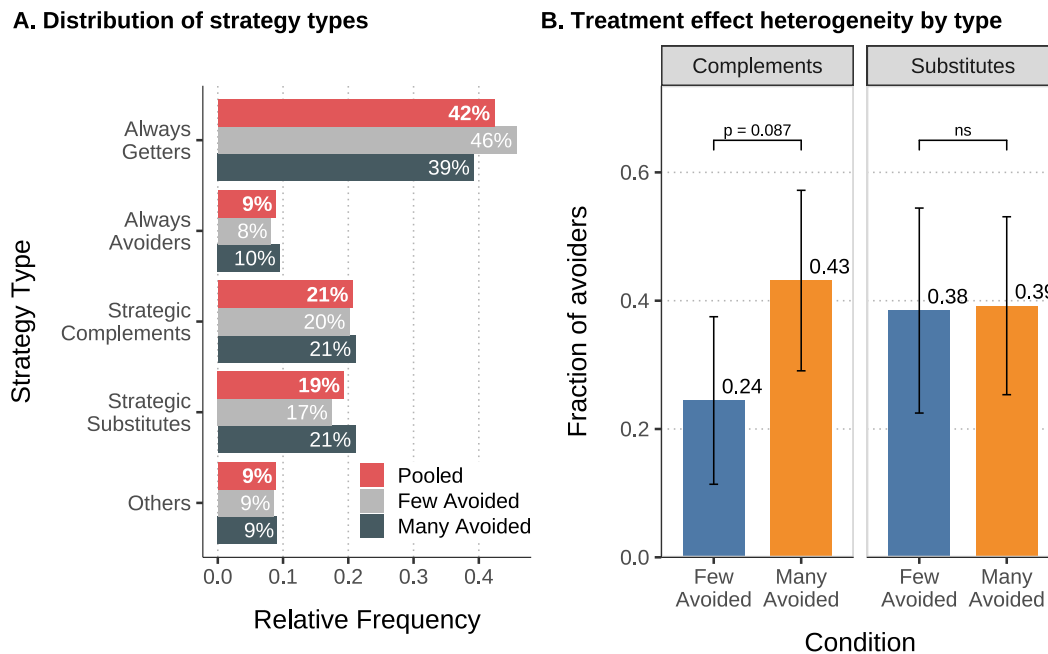
Note: The table reports estimates from a linear probability model regressing information avoidance on the treatment dummy. The dependent variable is a binary indicator equal to one if the subject avoid information (chose Later) and zero otherwise. The treatment dummy equals one if the subject was in the *Many Avoided* condition and zero in the *Few Avoided* condition. The regressions include strata fixed effects, demographic controls, field of study, and measures of economic and information preferences (see notes in Table 1.1). The number of observations drops in column (3) to (5) since five participants did not provide information about their age.

5.3 Heterogeneity of interdependence

Section 5.2 showed that exogenously increasing expectations about the share of avoiders in a group raises the likelihood of information avoidance, on average. A key question, however, is whether individual responses to the information decisions of others are heterogeneous. As Proposition 1 showed, theory allows for a variety of best-response types. To empirically map this heterogeneity, we employ the strategy method to elicit a *schedule* of best responses conditional on different shares of avoiders in the group. As detailed in Section 4.1, we categorize these empirical schedules based on whether they are increasing in the share of avoiders (Strategic Complements), decreasing (Strategic Substitutes), or constant (either Always Getters or Always Avoiders). This allows addressing the empirical question: do we observe the theoretical heterogeneity in practice? How large is this heterogeneity?

The data reveal substantial heterogeneity in how individuals respond to the information avoidance decisions of others. Panel A of Figure 1.6 presents the empirical distribution of individual strategy types in the whole sample (red bars). The most prevalent type is the Always Getter (42.4%), who seeks information regardless of the share of avoiding peers. At the opposite extreme, the least common type is the Always Avoider (8.8%), who avoids learning the state irrespective of the share of avoiders.

Figure 1.6: Heterogeneity of strategy types



Note: The figure illustrates the heterogeneity of strategy types. **Panel A** shows the distribution of types in the pooled sample (red bars), and separately for each treatment condition (gray bars), capturing the extensive margin (i.e., whether the treatment affects the distribution of types). **Panel B** shows the intensive margin: the treatment effect on the avoidance rate within each conditional strategy type (Complements and Substitutes). The y-axis represents the fraction of subjects who avoided information (chose Later). Error bars represent 95% confidence intervals; p-values derived from a two-sample proportion tests.

Beyond these unconditional types, we document the existence of conditional —or interdependent—

Table 1.3: Treatment effect on distribution of strategy types (extensive margin)

	Dependent variable: Subject belongs to strategy type			
	(1) Always Getter	(2) Always Avoider	(3) Complement	(4) Substitute
Many Avoided	-0.065 (0.046)	0.014 (0.026)	0.009 (0.038)	0.036 (0.037)
Control group mean	0.457*** (0.034)	0.081*** (0.018)	0.202*** (0.027)	0.175*** (0.026)
Observations	465	465	465	465

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table examines whether the treatment impacts the distribution of strategy types. It reports estimated coefficients from linear regressions where each column corresponds to a different strategy type. The dependent variable is a dummy equal to 1 if the subject's best-response schedule matches the given type and 0 otherwise. The independent variable is the treatment dummy (equal to 1 for the *Many Avoided* condition and 0 for the *Few Avoided* condition).

types, whose behavior depends on that of others. Among these types (approximately 40% of our sample), we identify two patterns. 20.6% of participants are Strategic Complements: they avoid information only if a sufficiently high share of their group members also avoids it. The proportion of Strategic Complements is statistically different from zero, with a 99% confidence interval of (0.171, 0.246). A similar share (19.4%) are Strategic Substitutes, who avoid information only if a sufficiently *low* share of their group members avoids it. The remaining minority (8.8%) are subjects who switch their information decisions multiple times and, hence, whose best responses oscillate. In sum, we find substantial heterogeneity in strategy types, with a significant share of individuals exhibiting dependence on the choices of others.

We now assess whether the treatment effects are heterogeneous across strategy types, examining both the extensive margin (whether the treatment affects the distribution of types) and the intensive margin (the probability of avoidance conditional on the type).

On the extensive margin, we find, reassuringly, that the treatment has no statistically significant effect on the distribution of types. Panel A of Figure 1.6 shows the distribution split by treatment condition (gray bars), and Table 1.3 reports the results of regressions where a dummy for each type is regressed on the treatment indicator. The treatment coefficient captures the change in the proportion of a type between *Many Avoided* and *Few Avoided*. We do not find statistically significant differences for any type.

The distributional invariance across treatments is consistent with theory. Best-response *schedules* themselves should not change unless some (perceived) parameter of the environment—beyond the expected behavior of others—changes. Since the treatment was designed to change only the expectation about others' behavior, but no other feature of the environment, there is no theoretical reason to expect the distribution of types to shift across treatments. Accordingly, we find that, empirically, the treatment has no effect on the extensive margin.

We now turn to the intensive margin. Panel B of Figure 1.6 compares the share of avoiders (who chose Later) between treatment conditions conditional on the strategy type. Table 1.4

Table 1.4: Treatment effect on information avoidance, by strategy type (intensive margin)

	Dependent variable: Avoided information			
	(1) Always Getters	(2) Always Avoiders	(3) Complements	(4) Substitutes
Many Avoided	0.014 (0.033)	0.056 (0.057)	0.187* (0.096)	0.008 (0.106)
Control group mean	0.049** (0.022)	0.944*** (0.057)	0.244*** (0.066)	0.385*** (0.080)
Observations	197	41	96	90

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table presents estimates from linear probability models where the dependent variable is an indicator equal to 1 if the subject avoided information (chose Later). Each column reports a separate regression estimated on the subsample of a distinct strategy type. The independent variable is the treatment dummy (equal to 1 for the *Many Avoided* condition and 0 for the *Few Avoided* condition).

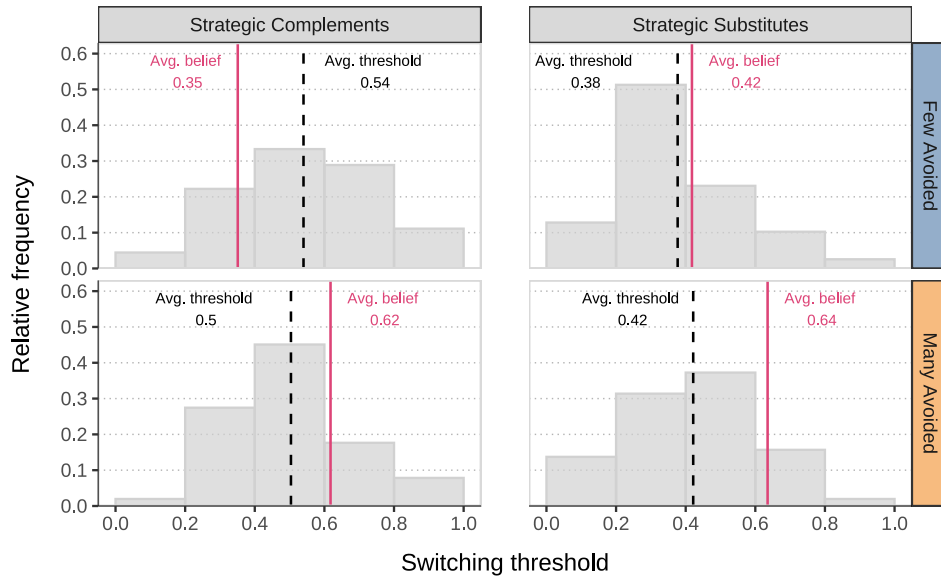
reports a linear regression of the avoidance dummy on the treatment for each type. In these regressions, the treatment coefficient captures a conditional average treatment effect (CATE): the difference in the likelihood of avoidance between the *Many Avoided* and *Few Avoided* conditional on the strategy type. We find that the *Many Avoided* treatment significantly increases the likelihood of information avoidance for Strategic Complements. This is in line with theoretical intuitions: since the treatment successfully raised beliefs about others' avoidance—the argument of best-response schedules—and Complements' schedules are increasing, they should respond by avoiding more often. By contrast, however, we found no significant response among Strategic Substitutes.

The difference in responsiveness between Strategic Complements and Substitutes can be explained by the interaction between switching threshold distributions and the treatment-induced shift in beliefs, shown in Figure 1.7. Average beliefs are shown in pink lines and average switching threshold in black lines. While the treatment leaves the average switching thresholds unchanged for both types (around 0.5 for Complements and around 0.4 for Substitutes), it substantially increases average beliefs (from 0.35 to 0.62 for Complements and from 0.42 to 0.64 for Substitutes). For Strategic Complements, the belief shift crosses the threshold, triggering increased avoidance. For Strategic Substitutes, however, the new beliefs remain on the same side of the threshold, resulting in no significant change in average avoidance rates.

Finally, columns (1) and (2) of Table 1.4 show that, as expected, the treatment does not affect the avoidance rate for Always Getters and Always Avoiders, who, by definition, follow strategies that are unresponsive to the behavior of others. Taken together, the pattern of intensive margins of treatment effects by types suggests that the positive average treatment effect documented in Section 5.2 is primarily driven by the responsive behavior of Strategic Complements.

Our analysis reveals considerable heterogeneity underlying the average treatment effect. The significant increase in information avoidance is not uniform, but is instead almost entirely

Figure 1.7: Switching thresholds and beliefs by strategy type and treatment condition



Note: The figure shows the switching thresholds—the fraction of other group members avoiding information at which a subject changes her own decision—along with beliefs about the fraction of others avoiding information, by strategy type and treatment condition. The distribution of individual switching thresholds is represented by gray bars. The average threshold is indicated by a black line, and the average belief about the share of avoiders is shown by a pink line.

driven by Strategic Complements, while the behavior of Always Getters, Always Avoiders, and Strategic Substitutes is unchanged. This pattern of responses is explained by the interaction between switching thresholds and beliefs.³⁰ Our findings suggest that understanding *which* individuals respond and *how* they respond (in terms of direction) is important for predicting aggregate acquisition of information. We explore this point through simulations that vary the composition of the group in terms of strategy types and trace the equilibrium prevalence of information avoidance in Section 6.

5.4 Mechanisms

Section 5.2 showed a positive average treatment effect. In this subsection we explore underlying mechanisms: what makes subjects in the *Many Avoided* condition avoid discovering the state more often than subjects in the *Few Avoided* condition?

The strategic complementarity in information avoidance could be driven by a variety of mechanisms. First, we focus on the “Mutually Assured Ignorance” (MAI) channel, formalized by Bénabou (2013) and central in our model: the idea that when a subject expects others to remain ignorant about the state, her own payoffs in the bad state become worse, making the event of discovering bad news about the state more aversive. We provide evidence supporting

³⁰Although Strategic Substitutes were unresponsive in this specific intervention, their behavior can potentially change under larger changes in beliefs. The data from the strategy method, which maps the best-response schedule for a wider range of beliefs, allow predicting decisions under such counterfactual beliefs.

this mechanism in Section 5.4.1 by using our KP measures of expected anticipatory utilities conditional on finding out good, bad, and no news. We then discuss the plausibility of alternative mechanisms in Section 5.4.2.

5.4.1 Anticipatory utility and the Mutually Assured Ignorance mechanism

Anticipatory utility and information decisions To shed light on determinants of information decisions and on mediators of the treatment effect, we use the KP questions to measure the expected anticipatory utility associated with the decisions of acquiring and avoiding information. First, we show that our measures of anticipatory utility significantly predict actual information decisions in our experiment, above and beyond the predicting power of the IPS, an established measure of information preferences. Next, we provide suggestive evidence that the treatment operates mainly through the utilities as captured by the KP measures. Finally, and crucially, we show that the treatment increases the net incentives of avoiding information by lowering the expected anticipatory utility of finding out *bad* news, but has no effect on the expected anticipatory utility of finding out *good* nor *no* news, in line with the MAI mechanism in Bénabou (2013) and with our model.

Recall that the KP questions elicit subjects' estimation of their own anticipatory utility *conditional on each possible information set*—that is, upon finding out good news ($\hat{v}_i(g)$), bad news ($\hat{v}_i(b + \Delta)$), and no news ($\hat{v}_i(\hat{p}_i b + (1 - \hat{p}_i)g)$)—. In addition, the KP questions elicit the subjective belief that the state is bad ($\hat{p}_i$). All the KP questions are elicited before subjects make their information decisions, so that we interpret them as *expected* anticipatory utilities, where expectations are taken before the subject potentially learns any information.

Based on the answers to the KP questions, for each subject we infer the expected anticipatory utility of acquiring information, by taking the expectation over states

$$U_{0,i}^{\text{Get}} \equiv \hat{p}_i \cdot \hat{v}_i(b + \Delta) + (1 - \hat{p}_i) \cdot \hat{v}_i(g) \quad (4)$$

based on the subjective belief about the state, $\hat{p}_i$. On the other hand, the expected anticipatory utility of avoiding information is simply

$$U_{0,i}^{\text{Avoid}} \equiv \hat{v}_i(\hat{p}_i b + (1 - \hat{p}_i)g) \quad (5)$$

We then estimate the net expected anticipatory utility of avoiding information φ_i , defined in Equation (2), as

$$\hat{\varphi}_i \equiv U_{0,i}^{\text{Avoid}} - U_{0,i}^{\text{Get}}, \quad (6)$$

The “hat” notation in $\hat{\varphi}_i$ emphasizes that this is our empirical estimate of the theoretical construct φ_i .

The model predicts that an individual i should avoid information whenever the net utility $\varphi_i > 0$, acquire it if $\varphi_i < 0$, and be indifferent if $\varphi_i = 0$. To assess the predictive power of our empirical anticipatory utility measures, we regress the avoidance dummy α_{ig} on the net

expected anticipatory utility, $\hat{\phi}_i$, or alternatively, a dummy variable indicating $\hat{\phi}_i > 0$.³¹ To benchmark our measures against an established alternative, we also include the IPS in the regressions. All specifications control for field of study and strata fixed effects, demographic characteristics, and economic preferences.

Panel A in Table 1.5 reports the results. Column (1) shows that, without the IPS or the KP measures, the controls explain only 6% of the variation in choices, as indicated by the adjusted R^2 . Column (2) adds the Information Preferences Scale (IPS). The IPS is a significant negative predictor of avoidance, consistent with the interpretation that individuals with a general willingness to *know* potentially undesirable information are *less* likely to avoid information in our experiment. Adding the IPS to the model improves explanatory power only modestly ($R^2 = 7\%$). Columns (3) and (4) replace the IPS with our measures of anticipatory utility. Both versions of anticipatory utility, continuous and binary, are highly significant (1% level) and positively associated with avoidance: a higher (positive) net anticipatory utility of avoidance is associated with a higher probability of avoiding information. Notably, models including our measures explain approximately 20% of the variation, nearly three times than that with the IPS. This is possibly because the KP questions elicit aspects that are specific to the decision environment. Columns (5) to (6) include both the IPS and KP measures. Perhaps surprisingly, for both measures, the coefficients remain remarkably stable in sign and magnitude. The KP measures also explain variance above and beyond the IPS. An interpretation is that the information decisions are driven by both a general disposition (captured by the IPS) and environment-specific aspects (captured by the KP measures). This pattern of findings supports both the predictive and discriminant validity of our KP measures, suggesting that the latter are not merely proxies of general preferences for information, but instead elicit aspects of preferences for information that are specific to our environment and predictive of decisions. Overall, our net utility measure is a more powerful predictor of information avoidance, and its predictive power is independent of that of the IPS.

Next, we assess whether anticipatory utilities mediate the treatment effect. Existing methodologies of causal mediation typically rely on the often strong assumption that the mediator is independent of unobservables conditional on the treatment. Since this assumption may not hold, the evidence reported in the remainder of this paragraph may be interpreted as tentative. Panel B presents regressions that include both the treatment dummy and the IPS or KP measures. For reference, column (1) reproduces the treatment effect estimate in the baseline specification from in Table 1.2. Column (2) adds to this regression the IPS as a benchmark, with the treatment coefficient remaining virtually unchanged. Columns (3) and (4) replace the IPS with our measures of net expected anticipatory utility. When these utility measures are included, the treatment effect shrinks substantially in magnitude and becomes statistically insignificant (for the continuous utility specification) or less significant (binary specification). Moreover, comparing the coefficients of the IPS and KP measures (columns 2-6) between Panel A (which exclude the treatment dummy) and Panel B shows that the inclusion of the treatment

³¹The dummy is equal to 0 if $\hat{\phi}_i < 0$, and encoded as a missing value for border cases where $\hat{\phi}_i = 0$ (12 individuals in our sample).

Table 1.5: Anticipatory utility and information avoidance

	Dependent variable: Avoided information					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Predictive power						
Information Preferences Scale (IPS)		-0.129** (0.052)			-0.136*** (0.047)	-0.148*** (0.049)
Net utility of avoidance (continuous)			0.023*** (0.003)		0.023*** (0.003)	
Net utility of avoidance (dichotomous)				0.372*** (0.046)		0.379*** (0.045)
Controls and fixed effects	✓	✓	✓	✓	✓	✓
Observations	460	460	460	448	460	448
Adjusted R ²	0.059	0.071	0.185	0.211	0.199	0.226
Panel B: Treatment and utility						
Many Avoided	0.122** (0.049)	0.123** (0.049)	0.061 (0.047)	0.081* (0.045)	0.062 (0.046)	0.080* (0.045)
Information Preferences Scale (IPS)		-0.130** (0.052)			-0.136*** (0.047)	-0.147*** (0.048)
Net utility of avoidance (continuous)			0.022*** (0.003)		0.022*** (0.003)	
Net utility of avoidance (dichotomous)				0.365*** (0.047)		0.372*** (0.046)
Controls and fixed effects	✓	✓	✓	✓	✓	✓
Observations	460	460	460	448	460	448
Adjusted R ²	0.070	0.082	0.187	0.214	0.200	0.230

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note: The table presents OLS estimates analyzing the predictive power and potential mediating role of the KP measures on actual information decisions. The dependent variable is an indicator equal to 1 if the subject chose to avoid information about the state (option Later) and 0 otherwise. Panel A assesses the predictive power of a number of measures. “Information Preferences Scale (IPS)” (Ho et al., 2021) measures overall preference to know potentially undesirable information; higher values indicate stronger preference to know. “Net utility of avoidance (continuous)” is the empirical measure of the net expected utility of avoiding information (relative to acquiring it), $\hat{\phi}^i$, derived from the KP questions. “Net utility of avoidance (dichotomous)” is an indicator equal to 1 if the empirical net utility of is higher than zero, $\hat{\phi}^i > 0$. In Panel A, coefficients are standardized for comparability across measures. Panel B focuses on how the treatment dummy (“Many Avoided”) changes upon inclusion of the predictive measures; coefficients are not standardized for comparability with Table 1.2. All regressions control for demographic characteristics (age and gender), economic and social preferences (patience, risk-seeking, trust, altruism, and positive and negative reciprocity), and fixed effects by strata and field of study of the participant.

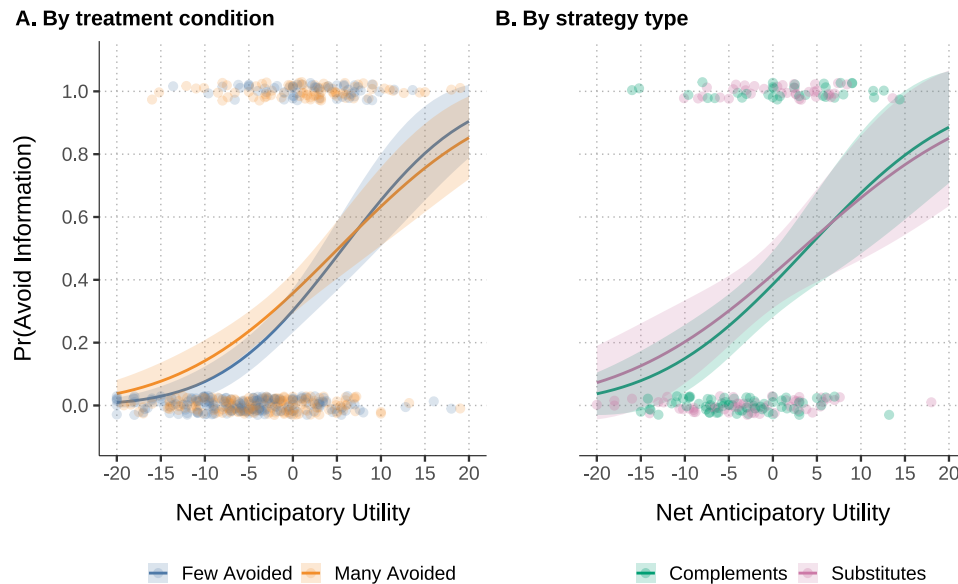
dummy leaves those coefficients essentially unchanged. This pattern of findings is consistent

with a mediating role for anticipatory utility: the treatment appears to operate primarily by influencing this utility. This is theoretically intuitive: the decision should depend only on the associated utility, so once we control for the utility, variations in the treatment dummy should become insignificant.

To further illustrate this point, in Figure 1.8 we report a probit model where the latent variable is the net utility $\hat{\phi}_i$. The model is estimated separately for the treated (*Many Avoided*) and control (*Few Avoided*) subsamples. The probability of avoiding information increases with the empirical net utility $\hat{\phi}_i$. Crucially, this relationship is statistically indistinguishable between treatment conditions. We repeat the same exercise across interdependent strategy types (Complements and Substitutes, where there is variation in choices) and find the same pattern. Again, this is in line with the theoretical notion that for predicting information avoidance, once we condition on the net utility, it shouldn't matter whether the subject is a Strategic Complement or Substitute.

Overall, these findings suggest that the KP measures, designed to measure anticipatory utility, proxy relevant determinants of information decisions and act as (or at least, are correlated with) a key mediator of the treatment effect.

Figure 1.8: Association between net expected anticipatory utility and information avoidance



Note: The figure illustrates the relationship between our empirical measure of the net expected anticipatory utility of avoiding information and the propensity to avoid information. The empirical net utility is calculated as the difference between the utility of avoiding information and the expected utility of acquiring information, elicited from the KP questions. Each point represents an individual participant's data, jittered vertically. The solid line shows the predicted probability of information avoidance from a probit regression model, estimated separately for each treatment condition (panel A) and each strategy type (panel B). The shaded bands indicate the 95% confidence intervals.

Prevalence of avoidance, anticipatory utility, and the contagion of avoidance Having established the validity of our KP measures, we now look deeper at specific components of the net utility of avoiding information. This allows us to provide a decomposition of the treatment

effect on such utility, useful to assess the plausibility of the MAI mechanism.

In separate regressions, we estimate the effect of the treatment on the following potential mediators: a) the net utility of avoiding information, $\hat{\phi}_i$, and b) each of its components separately; both excluding the set of controls and fixed effects (“minimal” models) or including them (“full” models). The treatment coefficient of each regression, capturing the effect on each mediator, is shown in Figure 1.9. These coefficients have a causal interpretation, and are standardized for comparability across measures. Starting from the top of the figure, the first pair of coefficients shows that the *Many Avoided* treatment has a positive and highly significant effect on the net utility of avoiding information—which, as shown above, is predictive of actual information choices. How is this effect decomposed? The second pair of coefficients shows the effect of the treatment on the expected utility of finding out *bad* news about the state upon acquiring information. The anticipatory utility of discovering bad news significantly deteriorates in the *Many Avoided* condition relative to the *Few Avoided* condition. This is consistent with the model, as the payoff in the bad state b (severity of the screams) worsens as more group members avoid information. In contrast, the third pair of coefficients shows that the treatment has no effect in the expected utility of finding out *good* news. This again aligns with the theoretical intuition: since the payoff of the good state (silence, g) is unaffected by the number of avoiders in the group, there should be no treatment effect on the utility of discovering good news. Finally, the treatment also increases the utility of avoiding information and learning *no* news.

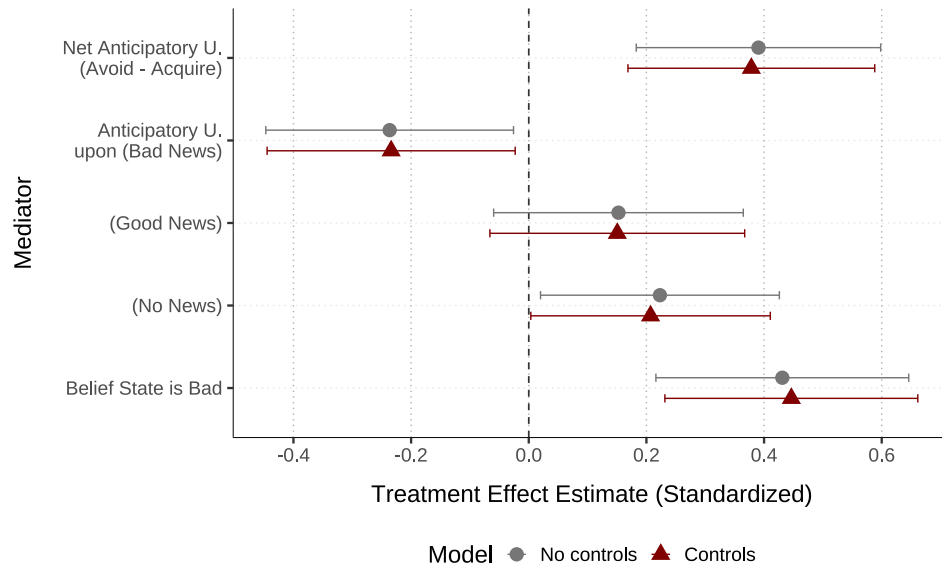
Unexpectedly, the treatment also increases the subjective belief that the state is bad. In principle, this suggests the possibility that changes in beliefs about the state in response to the treatment could be part of the mechanism, either as a competing channel or by interacting with and amplifying the MAI mechanism, as follows: updating beliefs about the information decisions of others may deteriorate the utility of learning bad news and also induce the subject to believe that the bad state is more likely, further reducing incentives to acquire information. The plausibility of this alternative channel can be informed by the empirical literature on the relationship between prior beliefs about the state and information decisions. In the setting closest to ours, Falk & Zimmermann (2024) find that, in an individual decision-making setting, experimentally manipulating prior beliefs about the bad state (p_i) does not lead to significant changes in information avoidance.

Taken together, the treatment effects on each utility component combine to increase the implied *net* anticipatory utility of avoiding information, suggesting that one of the main sources of increased incentives to avoid information arises from the downfall of anticipatory utility of finding out bad news, a result consistent with the MAI mechanism of Bénabou (2013) and of our model.

5.4.2 Other mechanisms

While the evidence is consistent with the MAI channel, the interdependence of information avoidance could, in principle, arise from other behavioral mechanisms. We discuss some of the main possibilities below.

Figure 1.9: Mechanisms: treatment effects on components of net anticipatory utility



Note: The figure shows the treatment effect estimate on variables measuring anticipatory utilities. For each dependent variable (listed on the y-axis), we report standardized coefficients from regressions on the treatment dummy. We show estimates from two specifications: a “Minimal” model with no controls and a “Full” model that controls for age, gender, economic and social preferences, the Information Preferences Scale, and strata and field-of-study fixed effects. The first dependent variable is the Net Expected Anticipatory Utility (the difference in expected anticipatory utilities between avoiding and acquiring information, derived from the KP questions). This net utility is decomposed according to equation (2) by the remaining variables displayed on the y axis: the anticipatory utility upon discovering good news (upon acquiring information and discovering the Quiet state), bad news (discovering the Screams state), and no news (avoiding information); and the subjective belief that the state is Bad. Coefficients are standardized for comparability across measures. Whiskers indicate 95% confidence intervals.

Conformity and herding. A first alternative possibility is that the positive correlation between beliefs about others’ behavior and own behavior is driven by conformist preferences. We say that an individual is conformist if he derives direct utility from matching his actions to those of others. Alternatively, the positive correlation may be driven by information herding. Under this mechanism, subjects are unsure about which is the most convenient course of action (i.e., to avoid or acquire information) in our particular experimental setting; if they believe that the signal about others’ behavior in the past is more informative about the most convenient decision than their own priors, they may rationally follow the crowd and match their actions to that of others. While each mechanism is grounded in distinct psychological motivations, they all formally predict the same empirical pattern: an individual’s likelihood of avoidance increases with the perceived fraction of avoiders in their group. Given the similar empirical predictions, we label these mechanisms jointly as “conformity” for brevity. To assess them, we conduct four tests. Specifically, if the observed treatment effect is purely driven by conformity, then we would expect four observable implications. First, since the treatment effectively changes the beliefs about others’ information decisions, the utility of acquiring information and discovering the *good* state should differ across conditions. We focus on the event of acquiring information and discovering good news because material payoffs g are invariant in the share of avoiders, and hence, constant across conditions (unlike the payoffs in

the bad state), allowing a cleaner comparison. As Figure 1.9 shows, we don't find support for this implication of conformity. Second, under a pure conformity channel the belief about the share of avoiders should predict own avoidance *independently* of the subjective belief about the state. In contrast, in the groupthink channel, what others do is relevant to the extent that the subject believes that the state is bad. To test these predictions, we regress the dummy α_{ig} on both beliefs and their interaction, with results reported in Appendix Table A.9. We find that, contrary to the conformity channel, the interaction of both beliefs is highly significant, suggesting that the belief about others' behavior predicts own avoidance only if the subject believes that the probability of the bad state is high. Moreover, once the interaction is included, the coefficient on the belief about others' information decisions becomes insignificant. This pattern lends support to the Bénabou's (2013) mechanism rather than to conformity. Third, conformity, if effective, would be manifested in Strategic Complements only. Assuming that subjects derive utility from imitating behavior of people from *both* past and current sessions, we should observe a shift in the switching threshold of Strategic Complements caused by the treatment, given that the strategy method elicits choices conditional on the choices of group members in the current, but not past, session. As Figure 1.7 showed, we do not find evidence of different switching thresholds across conditions. Fourth, under a pure conformity channel the effect of the treatment should not interact with reciprocity. However, we find mild evidence of the interaction, as shown in Table 1.2. Taken together, the four tests do not provide significant support for either the conformity or the information herding channel, while in some instances reinforce the evidence in favor of the groupthink mechanism.

Social learning. A second alternative is social learning. Under this mechanism, subjects do not know the parameters of their preference over screams, and use the signal about behavior in the past to update their beliefs about their own preference parameters. As a result, this may generate differences in information avoidance across treatment conditions. Although we do not have a direct test for this channel, we expect this mechanism to be negligible.

Non-linear preferences over volume. Finally, a third alternative mechanism relates to differences in the perceived benefit of a reduction of the 2 volume points across conditions. If subjects perceive that a 2-point reduction is less valuable in the *Many Avoided* condition relative to the *Few Avoided* condition, they would have higher net incentives to avoid information. While this mechanism is possible in principle, casual experience suggests that an increase of 2 volume points in a computer (out of 100 points) generates only a subtly noticeable difference in perceived loudness. This suggests that the effect through this mechanism would be negligible, unless subjects exaggerate their expectations of changes in loudness both substantially and asymmetrically between conditions.

6 Group composition and equilibria

In Section 5.3, we showed that individuals' best-response schedules to others' avoidance vary substantially. This heterogeneity suggests that the *equilibrium* prevalence of information avoid-

ance depends on the group’s composition of strategy types. Formally, we define an “information equilibrium” as an equilibrium fraction of information avoiders in a group. That is, a share π^* of avoiders constitutes an information equilibrium if a proportion π^* of group members best-responds to a share π^* by avoiding information while the remaining proportion $1 - \pi^*$ best-responds by acquiring it. In this section, we explore how different group compositions map into the resulting set of information equilibria.

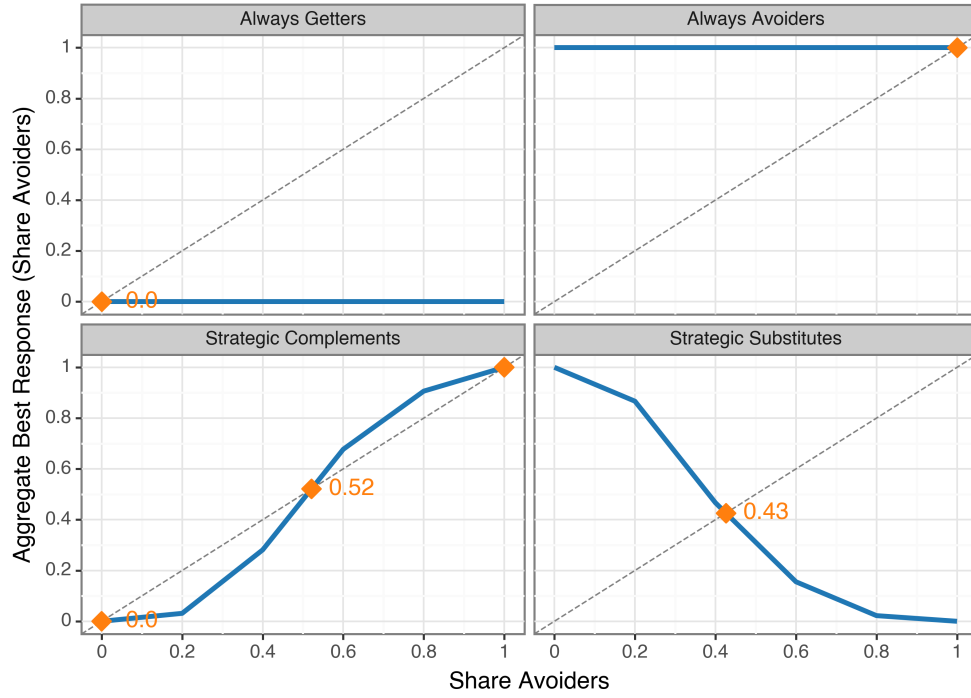
To derive the set of information equilibria of a group g , we first construct the group-level, or “aggregate”, best response function by aggregating the individual best-response functions of its members, as follows. The group is characterized by N_g individual best-response schedules, each defined as a function $f_{gi} : [0, 1] \rightarrow \{0, 1\}$ that maps each possible share of avoiders to a decision to avoid (1) or acquire (0) information. To derive the aggregate best response function F_g , we simply take the average of f_{gi} at each π , over group members. For example, consider a simple group composed by one Always Getter (with $f_{g1}(\pi) = 0$ for all π) and one Always Avoider (with $f_{g2}(\pi) = 1$ for all π). Then the aggregate best response function is $F_g(\pi) = 0.5$ for all π . In other words, for any proportion of avoiders π , half of the group members best-responds by avoiding information.

Because of implementation constraints, it is not possible to elicit f_i at *every* possible value of π : f_{gi} can be practically elicited for a coarse set of values of π . This yields a coarse representation of the group-level schedule F_g . To estimate its full shape, we impute the missing values by linear interpolation. This approach is simple, provides good local approximations around observed points of F_g under a continuity assumption, and ensures that the estimated schedule varies smoothly with π .

Once the full shape of F_g is estimated, identifying the information equilibria is straightforward: it is the set of fixed points of F_g , namely, the set of values π^* such that $F_g(\pi^*) = \pi^*$. Graphically, these are the points where F_g crosses the identity line.

To build intuition, we begin with “homogeneous” groups, namely, groups composed entirely by a single strategy type. Figure 1.10 plots the corresponding group best-response functions F_g by conditioning the empirical sample on each strategy type. Blue lines indicate the group’s best-response schedule. The set of equilibria, shown as orange diamonds, are given by the intersection(s) of F_g and the identity line (dotted). Homogeneous groups of unconditional types (Always Getters or Always Avoiders) display flat aggregate best-response functions, since members’ choices are unaffected by others’ decisions. Those groups have a unique information equilibrium at either extreme, where everybody is informed (Always Getters) or uninformed (Always Avoiders). In contrast, homogeneous groups of conditional types generate strictly monotonic F_g : negatively sloped for Strategic Substitutes and positively for Strategic Complements. Substitutes feature a unique interior equilibrium. Complements, however, exhibit multiple equilibria—including fully informed (MAA), fully uninformed (MAI), and an interior equilibrium—, consistent with the theoretical prediction of Bénabou (2013). This possibility of multiple of equilibria is important, as it captures the risk of collective ignorance, even when full awareness is also sustainable. The plot therefore illustrates the theoretical insights discussed in Section 3.1: Complements can sustain both the MAA and the MAI equilibrium, while Substitutes can reach none of those.

Figure 1.10: Aggregate best-response functions and information equilibria (homogeneous groups)



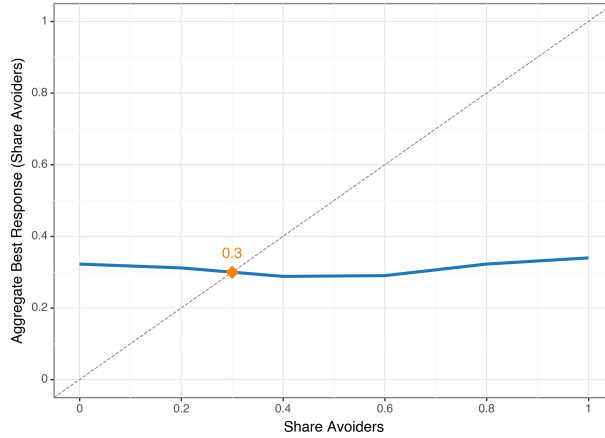
Note: The figure shows the aggregate (group-level) best response schedules of “homogeneous” groups composed by a single strategy type. Each group is formed by subsetting the experimental sample on the type indicated by the caption. Blue lines represent the group’s best-response schedule. The dotted, diagonal line shows the identity line. Orange diamonds indicate the fixed points of the schedule (i.e., the information equilibria).

Next, we consider how all strategy types are combined in the whole empirical sample. Figure 1.11 plots the estimated F_g for the pooled sample. Perhaps unexpectedly, the function is nearly flat, with a mild U-shape. The flatness of F_g arises from two sources. First, around half of the participants are unconditional types (Always Getters or Always Avoiders), contributing flat individual schedules f_{gi} ’s to the aggregation into F_g . Second, while 40% of subjects are conditional types, Complements and Substitutes are roughly balanced in number, so that, roughly speaking, their opposing slopes balance out each other. As a result, the aggregate F_g is nearly horizontal, with a unique interior equilibrium at $\pi^* = 0.30$.

Nonetheless, the diversity of shapes in aggregate reaction functions F_g shown in Figure 1.10 and the particular mix of types in our empirical sample highlights that the set of information equilibria may vary strongly with group composition. This motivates a more systematic analysis of how variations in composition influences equilibrium outcomes. To study this, we employ simulations based on bootstrap sampling: for each given composition, we generate random groups, compute their equilibrium sets, and trace how the resulting distribution of equilibria shifts as composition changes.

The detailed algorithm is described as follows. First, fix a group size N_g . Fix a sequence of group compositions $\tau \equiv \{\tau_t\}_{t=1,2,\dots,T}$, where each group composition τ_t is a vector

Figure 1.11: Aggregate best-response function and information equilibria (full empirical sample)



Note: The figure shows the aggregate (group-level) best response schedules of a group composed by full empirical sample. Blue lines represent the group's best-response schedule. The dotted, diagonal line shows the identity line. The orange diamond indicates the fixed point of the schedule (i.e., the information equilibrium).

$\tau_t \equiv (p_{AG}, p_{AA}, p_{SC}, p_{SS})_t$ whose elements are convex probabilities (respectively, for Always Getters, Always Avoiders, Complements, and Substitutes) that add up to 1. Start with $t = 1$, and the first composition τ_1 . For each bootstrap iteration $b = 1, \dots, B$, draw a random group of size N_g through bootstrap sampling with replacement, where the sampling probabilities of each strategy type given by τ_1 . This generates B simulated groups of size N_g with composition τ_1 . Among those B groups, for each group b , construct the aggregate best response F_b (including the linear interpolation) and find the set of fixed points Π_b , classifying their stability. Then, obtain composition-level statistics (computed across bootstrapped groups b): mean and 95% confidence interval of F_b at each π , and the distribution of information equilibria π^* . After this, let $t := t + 1$ and iterate for group composition τ_t until $t = T$.

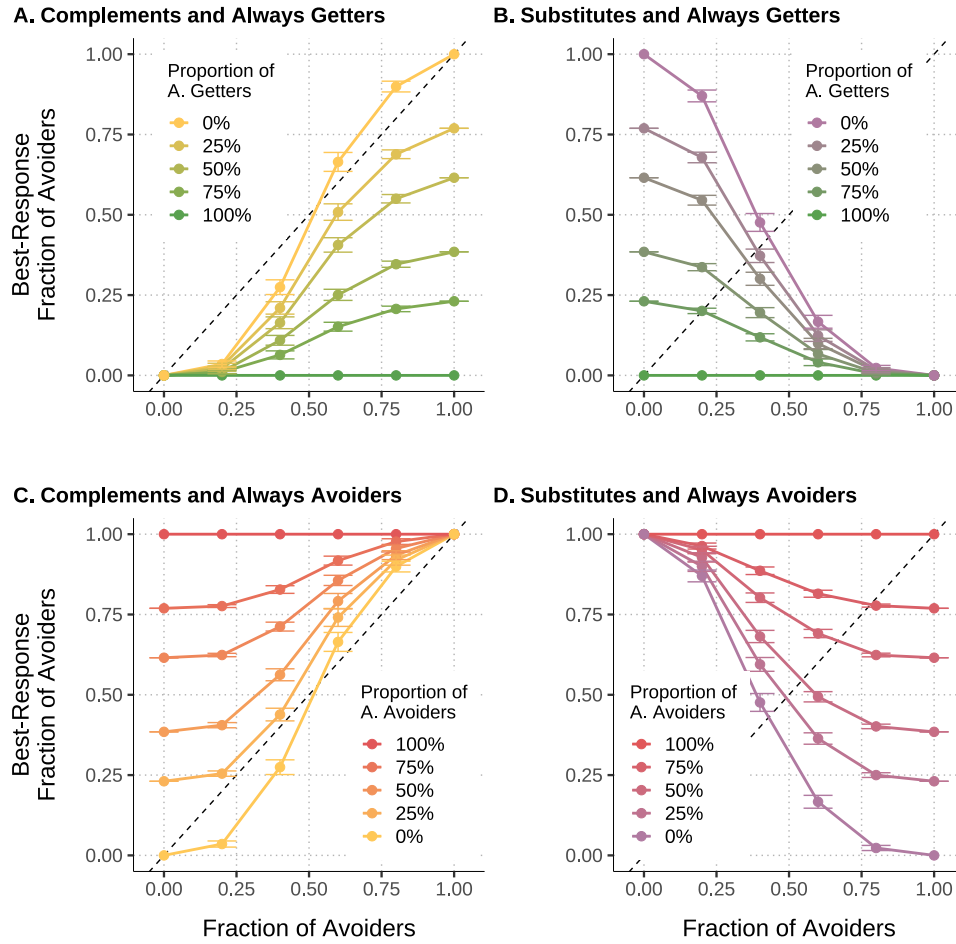
For the simulations, we iterate over a grid of all possible compositions subject to the constraint that the proportion of types take on discrete values from 0 to 1 in steps of 0.1. Given the additional constraint that the probabilities add up to 1, the grid includes 286 different group compositions. For each composition, we bootstrap $B = 100$ groups, each with $N_g = 13$ members.³² In principle, studying very large N_g would result in more precise estimates. However, for very large N_g it becomes unrealistic for a principal to be able to control the group composition. We therefore focus on finite group sizes, where composition remains policy-relevant. In total, the simulation involves 367,600 individuals across 28,600 groups, across 286 group compositions.

³²If N_g is a “round” number, such as 10, it is possible that, for some draws, continuous portions of F_b lie exactly on top of the identity line, leading to an infinite number of equilibrium points. This introduces technical details about how to count and adjust for these when calculating the relative frequency of equilibria, as well as apparent non-monotonicities in features of the distribution of equilibria as the group composition τ_t varies. To avoid these issues, we focus the discussion on cases constructed to prevent the emergence of infinite number of equilibria.

Restricted group compositions (two types) Among the many possible group compositions, we first focus on “restricted” sequences that begin with a group composed entirely of a conditional type and progressively add unconditional types. Formally, these satisfy $p_C + p_U = 1$, where $C \in \{SC, SS\}$ denotes a conditional type and $U \in \{AG, AA\}$ an unconditional type, and p_U progressively grows from 0 to 1.

Figure 1.12 shows how the average shape of the aggregate best-response function F_g evolves as the composition moves from fully conditional types ($p_U = 0$) to fully unconditional types ($p_U = 1$). The average is taken across the B simulated groups for each composition. The schedule of homogeneous groups of Complements ($p_{SC} = 1$) is plotted in yellow and that of Substitutes ($p_{SS} = 1$) in purple. The emerging pattern is that the introduction of unconditional types progressively flattens the schedule and pulls it towards the extremes. As unconditional types dominate, interior equilibria become less likely, and the multiplicity of equilibria gradually vanishes.

Figure 1.12: Effect of group composition on aggregate best-response schedules



Note: The figure shows aggregate best-response schedules under varying group compositions. Each panel focuses on mixtures of two strategy types. Lines show the average (across bootstrapped groups) aggregate best-response schedule for a given composition. Groups composed entirely of conditional types are shown in yellow (Strategic Complements) and purple (Strategic Substitutes). Within each panel, additional lines represent increasing shares of unconditional types—Always Getters (green) or Always Avoiders (red). Error bars denote 95% confidence intervals.

These tendencies are illustrated in Figure 1.13, which plots the distribution of fixed points (information equilibria) for different group compositions. When groups consist entirely of Strategic Complements ($p_{SC} = 1$), the distribution of equilibria is trimodal: bootstrapped groups exhibit full avoidance and full acquisition equilibria (both of which are typically stable), in addition to an interior equilibrium (typically unstable). As Complements are progressively mixed with Always Getters (Panel A), the distribution shifts to the left, so that the full-acquisition remains as an equilibrium, while the full-avoidance equilibrium gradually disappears. Conversely, mixing Complements with Always Avoiders (Panel C) shifts the mass toward the full-avoidance equilibrium and limits the full-acquisition outcome. In both cases, the interior equilibrium vanishes as the share of unconditional types increases, consistent with the flattening of F_g documented earlier.

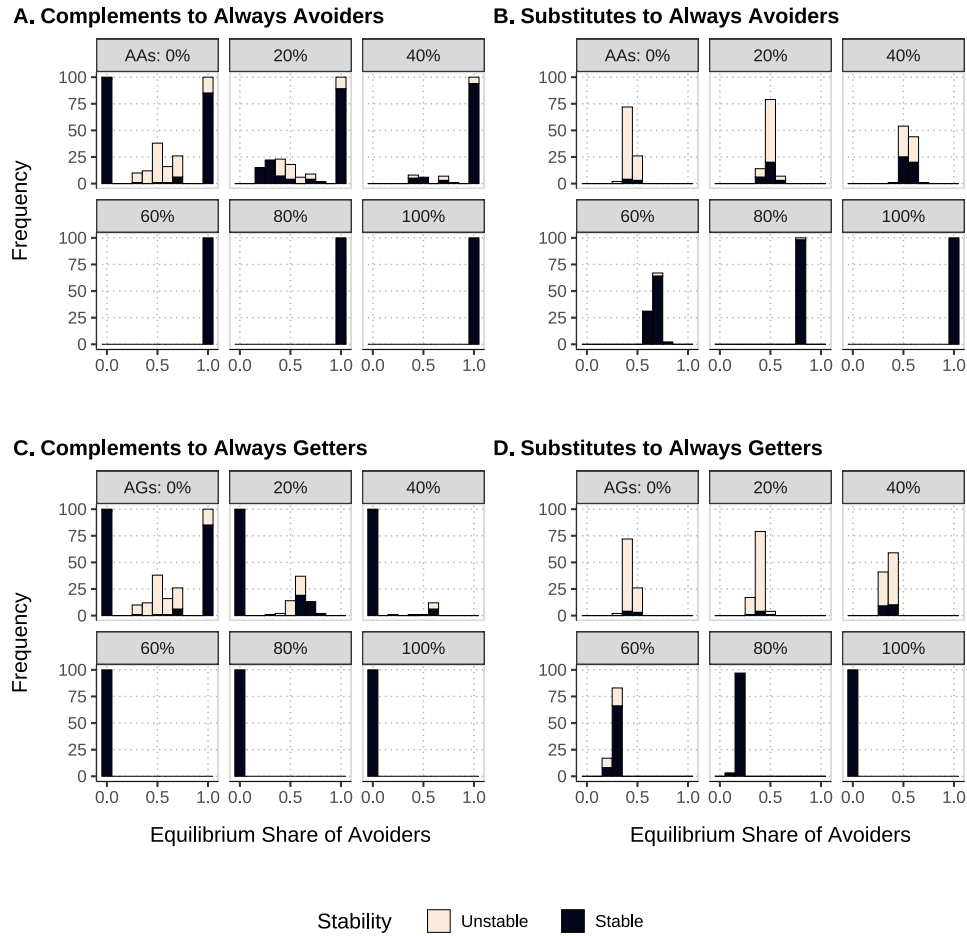
For Strategic Substitutes (Panels B and D), the picture is very different. Homogeneous groups of Substitutes ($p_{SS} = 1$) feature a single interior equilibrium. Notice however that these are typically unstable. Introducing Always Getters shifts this equilibrium leftward (toward fewer avoiders), while introducing Always Avoiders shifts it rightward (toward more avoiders). When the participation of unconditional types is sufficiently high, only corner equilibria remain (either full avoidance or full awareness).

Since, in practice, unstable equilibria are unlikely to persist and to be observed, in what follows we restrict attention to the set of stable equilibria.

While the histograms in Figure 1.13 provide a detailed view of the distribution of equilibria, they make it difficult to track how summary features of the distribution (such as the mean, median, and other quantiles) shift systematically with group composition. To capture these trends more transparently, Figure 1.14 plots the features of such distributions as the proportion of unconditional types increases. This representation is based on the same underlying distributions but highlights how the mean (orange diamonds), median (gray dots and lines), maximum equilibrium share of avoiders (red), minimum (blue), and other features evolve as compositions change.

The overall pattern is that (at least some) features of the distribution of equilibria respond monotonically but non-linearly to changes in the composition. For some statistics, such as the maximum and the minimum, there are regions in the composition sequence (x-axis) where the response (y-axis) is sharp, implying that the effect of a fixed amount of change in composition can vary greatly across the composition space. To illustrate, Panel A shows that introducing Always Getters (AG) into groups full of Complements can sharply impact some features, such as the maximum equilibrium share of avoiders or the 75% percentile (dark and light red lines). For example, once the share of AGs goes from 20% to 30%, the 75% percentile of the distribution of equilibria falls sharply from 0.6 to 0, meaning that, in 75% of the observed equilibria the share of avoiders is at most 60% (at 20% AGs), but falls sharply to 0% when only 10% additional AGs are introduced (at 30% AGs). Equivalently, having just 30% of AGs in a group of Complements ensures that the full-acquisition equilibrium is reached with 75% probability. Panel B shows that in groups of Substitutes, adding AGs shifts the features of the distribution of equilibria towards more acquisition rather smoothly. Panel C mirrors Panel A: adding Always Avoiders to Complements can strongly affect the lower percentiles of the distribution.

Figure 1.13: Group composition and the distribution of equilibrium shares of avoiders

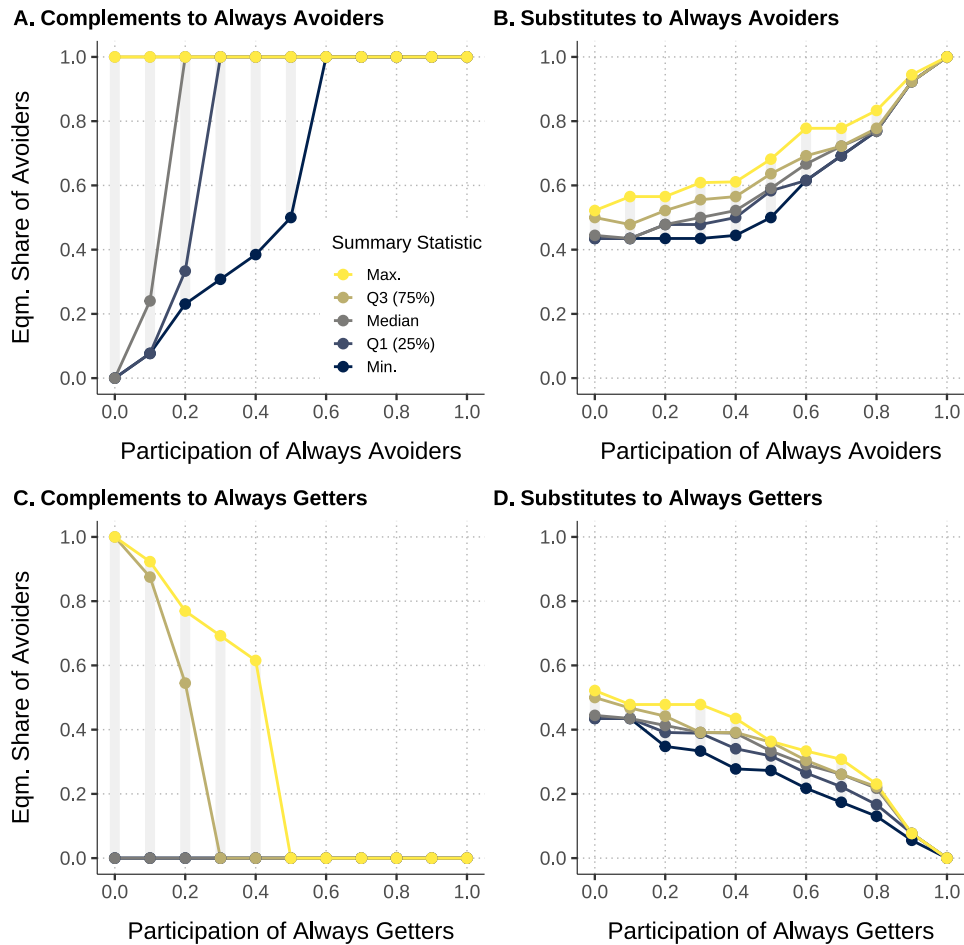


Note: The figure shows how the distribution of information equilibria (fixed points of the aggregate best-response schedules for each bootstrap group) varies with group composition. Each facet corresponds to a given share of unconditional types—Always Getters (AG, panels A and B) or Always Avoiders (AA, panels C and D)—with the remainder of the group composed of conditional types (Strategic Complements or Substitutes). Bars report the relative frequency (across bootstrap groups) of the equilibrium share of avoiders indicated on the x-axis. Stable equilibria are shown in blue, unstable equilibria in pink.

Overall, the figure suggests that groups of Complements are prone to tipping dynamics, while Substitutes seem to exhibit more gradual responses.

Higher-dimensional group compositions So far, we focused on restricted sequences of group compositions, where only two strategy types were mixed. We now turn to richer cases involving more than two types. To visualize these, it is useful to map the entire composition space onto a unit probability triangle. Since Substitutes are highly likely to produce unstable equilibria, we omit them in subsequent analyses. This allows redefining compositions as 3-dimensional probability vectors, which can be fully represented in the triangle. Figure 1.15 illustrates this representation. The proportion of Always Getters is represented on the x-axis;

Figure 1.14: Effect of group composition on distribution of information equilibria



Note: The figure illustrates the effect of gradual changes in group composition on the distribution of equilibrium avoidance rates. Each panel focuses on mixtures of two strategy types: Panel A replaces Strategic Complements with Always Getters, Panel B with Always Avoiders; Panel C replaces Strategic Substitutes with Always Getters, and Panel D with Always Avoiders. The x-axis represents the share of the unconditional type, while the y-axis reports the values of the equilibrium share of avoiders. Orange diamonds denote means across bootstrap samples within a group composition; vertical bars indicate interquartile ranges; and dots mark quantiles (max in red, 75% in pink, median in gray, 25% in light blue, and min in blue).

and Always Avoiders on the y-axis.³³ Each cell corresponds to a unique group composition, associated with a distribution of fixed points (with examples visualized in the histograms above).

A policy-maker may care about different features of such distributions, such as the mean, the median, or the variance. Panel A presents, for each cell, the average across all stable equilibria. A group of pure Complements (the cell at (0,0)) has an average close to 0.5, reflecting the near-symmetry of the fixed point distribution. As we increase the share of Always Getters (moving horizontally along the x-axis) the average equilibrium share of avoiders decreases

³³The proportion of Complements is implied by the complementary probability and visible as the horizontal (or vertical) distance from the cell to the hypotenuse of the triangle.

smoothly towards zero. Conversely, increasing the share of Always Avoiders (moving upward from (0,0) along the y-axis) pushes the average smoothly toward one. Overall, across the triangle, average equilibrium outcomes respond smoothly to group composition.

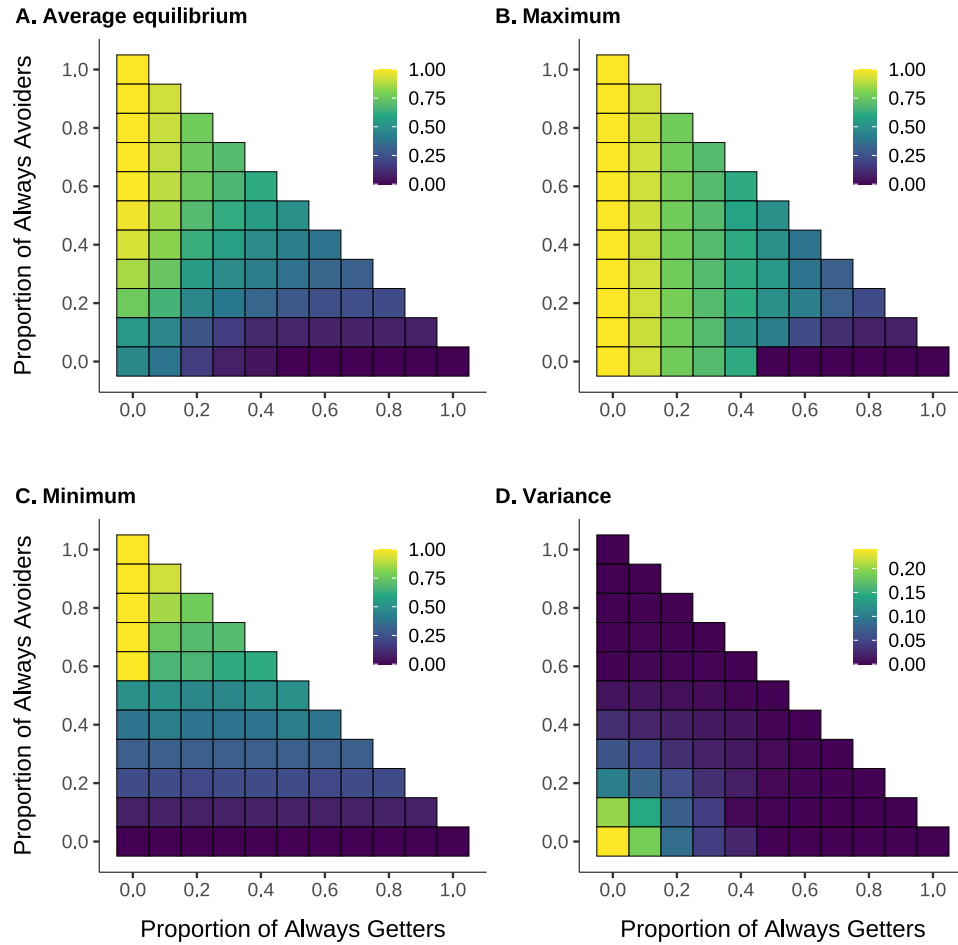
Panel B shows the maximum equilibrium in the distribution, for each cell. Intuitively, if a policy-maker regards avoidance of information as an undesirable behavior, panel B shows the “worst” equilibrium in the distribution, for each cell. For groups composed entirely of Complements, the worst equilibrium corresponds to full avoidance, and the addition of Always Avoiders does not change this. By contrast, introducing Always Getters improves the worse-case outcome: the maximum equilibrium share of avoiders declines monotonically. Notably, a strong discontinuity emerges around $p_{AG} = 0.5$. More generally, such discontinuities — where a fixed amount of change in composition produces large, abrupt shifts in the worst-case equilibrium — appear across the composition space, although these are stronger the higher the proportion of Complements. These discontinuities cluster around $p_{AG} = 0.5$: when the proportion of Always Getters is sufficiently high, the aggregate best-response function shifts downward enough to eliminate any stable interior equilibria, generating the jump in the maximum. The discontinuities weaken as the share of Always Avoiders increases, as their presence helps sustain interior equilibria.

Panel C shows the minimum among the equilibria (the “best” equilibrium in the distribution), with patterns analogous to those for the maximum. Here, the relevant probability is p_{AA} . Finally, panel D exhibits the variance of the equilibrium set, which may matter for a risk-averse policy-maker. Naturally, the variance is highest in groups consisting of pure Complements, where multiple equilibria exist. Adding unconditional types flattens the aggregate schedule, progressively collapsing the distribution of equilibria and hence reducing variance.

7 Conclusion

We live in an era of abundance of information in a variety of domains —politics, health, financial, environmental. Despite the wide availability of virtually free information, sometimes individuals prefer to avoid it, choosing, in effect, to remain in a state of willful ignorance. In this paper, we study whether this behavior is interdependent among individuals whose information decisions generate payoff externalities on other members. Our experimental evidence shows that, on average, willful ignorance can be contagious: individuals are more likely to avoid readily available information about the state when they expect others to do so, and hence, when they expect the outcome in the bad state to be more severe (and news revealing that the state is bad to be more aversive). Our findings suggest that the interdependence is mediated by an increase in expected anticipatory disutility of discovering bad news as a result of others’ information avoidance, and it is moderated by social preferences, particularly, by reciprocity. Beyond average effects, however, there is substantial heterogeneity in individual strategy types: while most individuals’ optimal information decisions are independent from others’ decisions, about 40% the subjects exhibit interdependence —they condition their information decisions on those of others. This heterogeneity of reaction functions suggests that the degree of aggregate contagiousness of willful ignorance may depend crucially on both the composition of the group

Figure 1.15: Group composition and features of equilibrium distributions



Note: The figure illustrates how group composition shapes key features of the distribution of equilibrium avoidance shares. Each cell in the unit probability triangle represents a distinct group composition, defined by the proportion of Always Getters (x-axis), Always Avoiders (y-axis), and Strategic Complements (implied by the vertical distance to the hypotenuse). Each panel reports a different feature of the equilibrium distribution—the average (A), maximum (B), minimum (C), or variance (D)—with values represented by the color gradient. Only stable equilibria are considered; Strategic Substitutes are excluded, as their equilibria are typically unstable.

in terms of reaction functions and the distribution of subjective expectations about others' behavior. Simulations show that, in effect, changes in group composition generate a wide variety of information equilibria, ranging from full information acquisition (and maximum mitigation of the bad state outcome) to full willful ignorance (and maximum severity).

Our study adds nuance to the recommendations from the literature on information preferences, which typically focus on individual decision settings. Our results suggest that incentivizing a person to acquire information (which may be beneficial for the individual in isolation) may generate positive or negative informational externalities to other members, in the form of stronger or weaker incentives to acquire information. Therefore, policy-makers should take into account these “cognitive externalities” and general equilibrium effects when designing policies that aim at increasing the aggregate take-up of information in groups when individuals' outcomes are interlinked.

Our results suggest some implications for organization design. Given that, usually, in organizations, information has instrumental value (i.e., subsequent decisions can be adjusted and improved if information about the state is acquired), a principal may be interested in incentivizing acquisition of information in her organization in order to minimize the emergence of a groupthink equilibrium where information avoidance is prevalent. Taking the group as given, a potential way to encourage information take-up is by correcting potential misperceptions about how others deal with information. Another way is to design the group composition in terms of strategy types when assembling teams. Our simulations show that the group composition has a strong impact in the distribution of equilibrium shares of informed members, and suggest the equilibrium responses can be non-linear in the mixture of types, in a way that there are regions in which the same amount of change in composition delivers the strongest responses.

Our investigation has some limitations and motivates avenues for future research. First, because our experimental setup is based on the setup of [Bénabou \(2013\)](#) where there is a theoretical link between agents' decisions and the severity of the outcome, our design cannot disentangle whether (and how much of) the contagiousness arises from variations in the severity of the outcome ("severity effect") or from variations in the proportion of others who engage choose to remain ignorant ("herding effect"). This limitation applies both to the average treatment effect and the individual best-response functions. We view our study as a first step in establishing the empirical existence of interdependence in information decisions. Further research is needed to disentangle the severity effect from the herding effect. Second, we consider a static decision-making setting, where all individuals choose their information decisions simultaneously. In many applications, such decisions occur sequentially, motivating future studies on willful ignorance in groups in dynamic settings.

Our study highlights that understanding the ways in which individuals react to others' information decisions is crucial for designing environments that promote the acquisition of information, which, usually, improves subsequent decision-making in organizations.

References

- Beaurenaut, M., Tokarski, E., Dezechache, G., & Grèzes, J. (2020). The ‘Threat of Scream’ paradigm: A tool for studying sustained physiological and subjective anxiety. *Scientific Reports*, 10(1), 12496.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2), 608–650.
- Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *The Review of Economic Studies*, 80(2), 429–462.
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.
- Eil, D. & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, 3(2), 114–138.
- Engelmann, J. B., Lebreton, M., Salem-Garcia, N. A., Schwardmann, P., & van der Weele, J. J. (2024). Anticipatory Anxiety and Wishful Thinking. *American Economic Review*, 114(4), 926–960.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, 133(4), 1645–1692.
- Falk, A. & Zimmermann, F. (2024). Attention and Dread: Experimental Evidence on Preferences for Information. *Management Science*, 70(10), 7090–7100.
- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3), 397–404.
- Ganguly, A. & Tasoff, J. (2017). Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future. *Management Science*, 63(12), 4037–4060.
- Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). LIONESS Lab: A free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.
- Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information Avoidance. *Journal of Economic Literature*, 55(1), 96–135.
- Haaland, I., Roth, C., & Wohlfart, J. (2023). Designing Information Provision Experiments. *Journal of Economic Literature*, 61(1), 3–40.
- Hertwig, R. & Engel, C., Eds. (2020). *Deliberate Ignorance: Choosing Not to Know*. Strüngmann Forum Reports. Cambridge, Massachusetts: The MIT Press.

- Ho, E. H., Hagmann, D., & Loewenstein, G. (2021). Measuring Information Preferences. *Management Science*, 67(1), 126–145.
- Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291.
- Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, 38(2), 95–115.
- Kreps, D. M. & Porteus, E. L. (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica*, 46(1), 185.
- Kumar, S., Forster, H., Bailey, P., & Griffiths, T. (2009). Mapping unpleasantness of sounds to their auditory representation. *The Journal of the Acoustical Society of America*, 124, 3810–7.
- Masatlioglu, Y., Orhun, Y., & Raymond, C. (2023). Intrinsic Information Preferences and Skewness. *American Economic Review*, 113(10), 2615–2644.
- Meissner, T. & Pfeiffer, P. (2022). Measuring preferences over the temporal resolution of consumption uncertainty. *Journal of Economic Theory*, 200, 105379.
- Momsen, K. & Ohndorf, M. (2023). Information avoidance: Self-image concerns, inattention, and ideology. *Journal of Economic Behavior & Organization*, 211, 386–400.
- Oprea, R. & Yuksel, S. (2022). Social Exchange of Motivated Beliefs. *Journal of the European Economic Association*, 20(2), 667–699.
- Oster, E., Shoulson, I., & Dorsey, E. R. (2013). Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease. *American Economic Review*, 103(2), 804–830.
- Pagel, M. (2018). A News-Utility Theory for Inattention and Delegation in Portfolio Choice. *Econometrica*, 86(2), 491–522.
- Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. *Beiträge zur experimentellen Wirtschaftsforschung*.
- Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy*, 69(3), 213–225.
- Suvorov, A., van de Ven, J., & Villeval, M. C. (2024). Selective Information Sharing and Group Delusion.
- Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98(5), 1829–1863.
- Tversky, A. & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.

CHAPTER 2

Confirmation-biased evaluation and labor market discrimination[†]

1 Introduction

Discrimination – less favorable treatment of members of a minority group with respect to otherwise identical members of a majority group – continues to be widespread. Systematic evidence of persistent discrimination is provided in [Rodgers \(2009\)](#) and, more recently, in [Bertrand & Duflo \(2017\)](#). Discriminatory treatment in the labor market, in particular, has been credited as the major cause behind the existing disparities among social groups ([Darity & Mason, 1998](#)) – for reviews of theory and empirical findings, see [Cain \(1986\)](#), [Altonji & Blank \(1999\)](#), and [Lang & Lehmann \(2012\)](#).

It is customary to classify theoretical models of employment discrimination according to the source of the discrimination. In taste- (or preference-) based theories, originating in the work of [Becker \(1957\)](#), employers suffer a disutility when interacting with members of a particular group. Models of statistical discrimination ([Phelps, 1972](#); [Arrow, 1973](#)), on the other hand, assume potential employers have no animus against any particular group, but cannot observe the productivity of individual group members, and thus use group identity to form beliefs about an employee’s productivity. For the most part, employers’ beliefs are taken as accurate, but there is a recent literature pointing to the importance of statistical discrimination based on inaccurate or incorrect beliefs (see [Bohren et al., 2025](#), for a discussion of the sources for inaccurate beliefs, and a review of the extant evidence).

This paper builds on the inaccurate statistical discrimination literature to examine patterns of discrimination when employers hold incorrect beliefs, and suffer from confirmation bias when evaluating employees from different social groups. Confirmation bias – the tendency ‘to misinterpret ambiguous evidence as confirming [one’s] current hypothesis about the world’ ([Rabin & Schrag, 1999](#)) – has been systematically studied by Social Psychology since the 1960s

[†]This chapter is based on joint work with Christian Ruzzier and was published in the *Journal of Economic Behavior and Organization*, June 2023, under the title “Discrimination with inaccurate beliefs and confirmation bias”. The published version is available at [this link](#).

and by Economics more recently.¹ Learning, in turn, is important in labor markets and within organizations, as it has implications for the wages, job assignments, promotions, and sector affiliations of individuals, and for changes in these variables over workers' careers (Gibbons et al., 2005).

There are several good reasons to expect confirmation bias to play an important role in our setting. Whether rooted in cognitive processes (specific cognitive limitations, lack of understanding of logic), or caused by motivational forces (maintain self-esteem and positive regard by others, protect one's ego, avoid certain type of errors, accomplish specific practical goals), and whatever the form confirmation bias takes, the evidence shows that, on the net, when people err, it tends to be in the direction of confirmation (see the extensive and widely cited reviews of Klayman, 1995 and Nickerson, 1998). In the labor market and within organizations, in particular, the specialized press and practitioners in general seem to regard confirmation bias as an important issue, especially in hiring. Wright & Schepker (2015), for instance, blame confirmation bias for mistakes in choosing C-suite candidates. Kahneman et al. (2019) argue further that the same bias is present in unstructured decision-making in general (whether in job interviews or in other, more strategic decisions).²

On the other hand, performance evaluations are in many instances plagued by ambiguity and subjectivity, and evidence is open for interpretation. The presence of ambiguous evidence or behavior that must be interpreted is fertile ground for confirmation bias, as is widely acknowledged in the literature (Keren, 1987, Griffin & Tversky, 1992, Klayman, 1995, Rabin & Schrag, 1999). Moreover, research shows that in these contexts, when people perceive themselves as objective evaluators, their judgments tend to be relatively more influenced by stereotypic beliefs – and that, however, many organizational contexts seem to encourage a sense of personal objectivity (Uhlmann & Cohen, 2007). When faced with the overload of information associated with a high number of applicants, employers appear to rely heavily on stereotypes to reduce uncertainty and simplify their decisions, and confirmation bias plays a role in those decisions (Uhlmann & Silberzahn, 2014).

We develop a simple model of wage formation in which potential employers (firms) start with incorrect beliefs about the relationship between group membership (e.g., race or gender) and productivity – that is, firms hold inaccurate stereotypes (cf. Altonji & Pierret, 2001) about workers. Information about workers' productivity is publicly learned over time. Learning is thus symmetric, but when firms update their beliefs about a worker's productivity to set wages, they suffer from confirmation bias, so that firms tend to interpret signals about productivity as if they were closer to their pre-existing beliefs.

In this setup, we show that firms that start with incorrect beliefs do not correct them fully,

¹Some have gone as far as to state that confirmation bias is “perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning” (Evans, 1989, p. 41), and that “if one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration” (Nickerson, 1998, p. 175).

²More systematic evidence is still in short supply (see Whysall, 2018 for a review of cognitive biases and discrimination in recruitment, selection, and promotion), and mainly concerns hiring decisions (see, e.g., Uhlmann & Cohen, 2007, and Uhlmann & Silberzahn, 2014, on gender discrimination in hiring).

leading to persistent wage discrimination in the labor market, i.e., different wages for otherwise identical workers belonging to different social groups. Put differently, with confirmation-biased learning, stereotypes have long-lasting effects on wages, whereas with Bayesian learning (as usually assumed) stereotypes would eventually be corrected – discrimination cannot persist in the long run in such a setup.

When workers only differ in group membership and potential employers hold a negative stereotype about the minority group, the minority worker is persistently discriminated against because her expected long-run wage is lower than the long-run wage of an otherwise identical majority worker. The magnitude of the discrimination will depend on the entrenchment of initial beliefs (how strongly held are the initial stereotypes and how willing are potential employers to change them). In the general case, when different groups have different perceived variabilities and the market interprets signals differently depending on the group identity of the worker, negative stereotypes generate discrimination against minority workers upon entry to the labor market, but are not enough to have discrimination in the long run, and reversals in discrimination are possible.

Our setup is amenable to thinking about whether interventions aimed at reducing discrimination would succeed if confirmation bias is an important source of discrimination. For instance, reducing stereotyping is predicted to reduce discrimination, consistent with evidence, but would only eliminate it in the long run in very special circumstances. Attracting employers' attention to within-group differences among minority workers or reducing noise, ambiguity or subjectivity in the evaluation of minority workers all have the potential to reduce discrimination (as some evidence suggests), but will achieve so only if minority workers are perceived as less productive than they truly are at the outset.

We then extend the model to a situation where employers in the same labor market are heterogeneous in terms of their initial beliefs (for example, employers who have the same group identity as a potential employee may have different initial beliefs about worker productivity from employers who do not), and show that the same set of factors that produce discrimination can also produce segregation. The model predicts that, in a cross section of occupations, after controlling for relevant differences across workers, discrimination against minority workers should be more pronounced in partially segregated occupations than in fully segregated ones. It also predicts, consistent with evidence, a negative correlation between the average pay (of all workers combined) and the proportion of all employees in that job that belong to the minority group.

We model confirmation bias as a source of persistent discrimination among individuals belonging to different social groups, over which society (and hence the market) might have different priors (or stereotypes). [Wason \(1960\)](#), [Lord et al. \(1979\)](#), and [Anderson et al. \(1980\)](#) are important early contributions on confirmation bias in Social Psychology. Within Economics, one of the first papers to address confirmation bias formally was [Rabin & Schrag \(1999\)](#). In their model, agents receive noisy signals about the state of the world. To model confirmation bias, they assume an agent misinterprets –with exogenous probability– a signal as supporting her current hypothesis. Confirmation bias is shown to lead to overconfidence in the favored hypothesis, and even an infinite amount of information may be insufficient to overcome its

effects: over time an agent may with positive probability come to believe with near certainty in the wrong hypothesis.

Where others have used [Rabin & Schrag \(1999\)](#) to introduce confirmation bias in their models (like [Müller, 2010](#), to study job rotation within a firm as a means to mitigating confirmation bias; [Pouget et al., 2017](#), to analyze the impact of confirmation bias on financial markets; or [Rabin & Schrag, 1999](#) themselves, to explore the consequences of confirmation bias for incentive provision in a principal-agent relationship), we adopt the model of confirmation bias and belief formation proposed by [Fryer et al. \(2019\)](#) –which formalizes the type of confirmation bias described by [Lord et al. \(1979\)](#) and [Darley & Gross \(1983\)](#). [Fryer et al. \(2019\)](#) introduce a human memory storage limitation as a foundation for the bias assumed in [Rabin & Schrag, 1999](#), and develop a model of confirmation bias for the case of continuous signals, which provides some new results showing that bias always occurs (with probability one) and depends on early signals and not just the prior.³

We use this model as a building block, and add it to a labor market setting with the aim of exploring the consequences of employers being subject to this cognitive bias on the wage gap. In addition, we enrich the market by introducing heterogeneity in prior beliefs (at least some of which will necessarily be inaccurate) about the productivity of different social groups, which allows us to characterize not only differences in wages but also patterns of labor market segregation.

The concept of confirmation bias has seen productive applications in several fields in economics in recent years, like industrial organization (e.g., [Mullainathan & Shleifer, 2005](#)), finance (e.g., [Pouget et al., 2017](#)), political economy (e.g., [Mullainathan & Washington, 2009](#)), marketing (e.g., [Narasimhan et al., 2005](#)), and organizational design (e.g., [Müller, 2010](#)). We contribute to this list by studying the consequences of confirmation bias for (wage) discrimination in the labor market.

The paper also contributes to a growing literature that has been recently expanding the classical models of statistical discrimination.⁴ Our work is closest to extensions of classical models that introduce behavioral elements, especially those considering inaccurate beliefs as a source of discrimination.⁵ [Barron et al. \(2025\)](#) differentiate between two different forms of belief-based discrimination, explicit (against women who are equally qualified than men) and implicit (against women who are differently qualified), and produce experimental evidence on these forms. [Bohren et al. \(2019\)](#) develop a model with evaluators who can have a misspecified model of inference about the distribution of ability conditional on gender, or about the preferences or beliefs of other evaluators. They show that such inaccurate beliefs may lead to discrimination based on gender, and enrich the traditional taxonomy of sources of discrimination by distinguishing between preferences (driving taste-based discrimination), rational beliefs (driving accurate statistical discrimination), and inaccurate beliefs (driving

³[Schwartzstein \(2014\)](#) provides a different foundation for confirmation bias, and also provides a discussion of how it can lead to persistent discrimination.

⁴These developments are aptly reviewed in [Onuchic \(2023\)](#).

⁵See, e.g., [Bordalo et al. \(2016\)](#); [Bohren et al. \(2019, 2025\)](#); [Coffman et al. \(2021, 2024a,b\)](#); [Campos-Mercade & Mengel \(2024\)](#); [Erkal et al. \(2022\)](#); [Lepage \(2024\)](#).

inaccurate statistical discrimination).

In [Bohren et al. \(2019\)](#), employers are Bayesian, except for a potential misspecification of their initial beliefs. Our paper departs from their setup by considering employers who, beyond any misspecification in their model of inference, exhibit non-Bayesian learning. This relates our paper to the broader literature on non-Bayesian updating (see the extensive review in [Benjamin \(2019\)](#)). On this front, [Campos-Mercade & Mengel \(2024\)](#) ask experimental subjects to hire one of two potential candidates with potentially different and unobservable productivity, and an observable education signal –and find evidence of substantial conservatism, and of excess discrimination (relative to a Bayesian benchmark), stemming from the neglect of workers’ education certificates. [Eyting \(2025\)](#) connects non-Bayesian learning and discrimination based on inaccurate beliefs. In particular, the author identifies the role of motives in generating inaccurate beliefs that lead to discriminatory outcomes. In an experimental setting, [Eyting \(2025\)](#) finds that subjects who play the role of employers exhibit asymmetric learning, weighing more the signals that are aligned with their motives.⁶

2 A model of learning with confirmation bias

We start from a simple dynamic model of learning and wage determination (see, e.g., [Freeman, 1977](#); [Harris & Holmstrom, 1982](#); [Farber & Gibbons, 1996](#)). Consider a competitive labor market in which a large number of identical, risk-neutral firms compete for the labor services of risk-neutral workers of initially unknown productivity $\eta \in \mathbb{R}$. Information is incomplete but symmetric: both workers and potential employers in the market share the initial belief that η is normally distributed with mean μ_1 and variance $1/h_1$ (where h_1 is the precision). Firms learn about each worker’s productivity by observing the worker’s output through time ($t = 1, 2, 3, \dots$), which is given by the following technology:

$$y_t = \eta + \varepsilon_t.$$

The stochastic noise term ε_t is assumed to be independently and identically distributed as $\varepsilon_t \sim \mathcal{N}(0, 1/h_\varepsilon)$ – hence, output is a noisy signal of a worker’s true productivity, and learning is gradual.

Workers have the following (publicly-known) utility function:

$$U(w) = \sum_{t=1}^{\infty} \beta^{t-1} w_t,$$

where $\beta \in (0, 1)$ is a discount factor, and w is a wage stream. Only short-term, non-contingent wages can be offered in this market, and we assume that wages are paid in advance. The wage offered at the beginning of any period t may, however, depend on the history of output

⁶We note that since, in the experiment, the employers’ motives are identified by their prior beliefs about the productivity distribution of each group (before an information treatment that corrects such beliefs), the results may also be consistent with confirmation-biased learning.

realizations up to time t , $y^t \equiv (y_1, y_2, \dots, y_t)$, through the updated beliefs about worker productivity – we denote such a wage by $w_t(y^{t-1})$.

The only decision workers make in this model is which firm they work for. Workers can leave firms in any period at no cost, and will work for the firm offering the highest current wage. Since the labor market is competitive, in each period wages are bid up to expected output (conditional on y^t), that is, until firms earn no profits:

$$w_t(y^{t-1}) = E[y_t | y^{t-1}] = E[\eta | y^{t-1}]. \quad (1)$$

In each period, firms make wage offers determined by their current beliefs about worker productivity; then each worker chooses for which firm to work, and finally individual outputs for that period are realized and observed by everyone.

To introduce confirmation bias into the learning process, we depart from traditional Bayesian models and consider each potential employer to be an interpretive evaluator as proposed by [Fryer et al. \(2019\)](#). An interpretive evaluator first interprets the signal given her prior, and then updates her beliefs following Bayes' rule, but using her interpretation rather than the raw information. This double updating amounts to weighing the prior belief twice and leads to confirmation bias.

Let $\hat{y}$ denote the interpretation of the signal y , and $\hat{\mu}$ the posterior mean based on the interpretation $\hat{y}$. The two steps involved in the learning process are the following:

1. INTERPRETATION OF THE SIGNAL

$$\hat{y}_{t+1} = \hat{\mu}_t \cdot \left(\frac{h_t^\mu}{h_{t+1}^y} \right) + y_t \cdot \left(\frac{h_\varepsilon}{h_{t+1}^y} \right),$$

where h_t^y and h_t^μ denote the precisions of the interpretation $\hat{y}_t$ and of the interpretive belief $\hat{\mu}_t$ at time t . The interpretive evaluator interprets (possibly ambiguous) information y_t based on her pre-existing belief $\hat{\mu}_t$, following Bayes' rule. This has the effect of 'pulling' the signal towards her pre-existing belief, and can be thought of as a model of the 'information assimilation bias' pointed out by [Lord et al. \(1979\)](#).

2. BELIEF UPDATING BASED ON THE INTERPRETATION

$$\hat{\mu}_{t+1} = \hat{\mu}_t \cdot \left(\frac{h_t^\mu}{h_{t+1}^\mu} \right) + \hat{y}_{t+1} \cdot \left(\frac{h_{t+1}^y}{h_{t+1}^\mu} \right).$$

Note that firms update their prior beliefs based on their *interpretation* $\hat{y}$.

Besides the double updating, the learning process is well known (see [DeGroot, 1970](#)). The posterior means $\hat{\mu}_{t+1}$ and precisions h_{t+1}^μ are given by:

$$\begin{aligned} \hat{\mu}_{t+1} &= \mu_1 \cdot \left(\frac{2^t h_1}{2^t (h_1 + h_\varepsilon) - h_\varepsilon} \right) + \left(\frac{2^t h_\varepsilon}{2^t (h_1 + h_\varepsilon) - h_\varepsilon} \right) \cdot \left[\sum_{j=1}^t y_j \left(\frac{1}{2} \right)^j \right], \\ h_{t+1}^\mu &= 2^t (h_1 + h_\varepsilon) - h_\varepsilon. \end{aligned}$$

See the online appendix for the derivation of these expressions and an analysis of the evolution of beliefs over time. From (1), the wage function is given by $w_t(y^{t-1}) = \hat{\mu}_t$.

The expected wage in an equilibrium with confirmation bias is:

$$E[w_t(y^{t-1})] = \mu_1 \cdot \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\epsilon) - h_\epsilon} \right) + \eta \cdot \left(\frac{2^{t-1} h_\epsilon}{2^{t-1} (h_1 + h_\epsilon) - h_\epsilon} \right) \left[\sum_{j=1}^{t-1} \left(\frac{1}{2} \right)^j \right].$$

As we can see, the confirmation-biased equilibrium wage is determined by the initial belief μ_1 to a larger extent than in the standard Bayesian case, where the weight attached to the prior mean would be $h_1 / (h_1 + (t-1) h_\epsilon)$.

In the long run, as the number of output realizations grows large ($t \rightarrow \infty$), the market's posterior mean, and thus wages, converge in expectation to something that is wrong with probability one, as in [Fryer et al. \(2019\)](#), as long as $h_1 > 0$ – a fairly general condition:

$$E[w_t(y^{t-1})] \xrightarrow[t \rightarrow \infty]{} \left(\frac{h_1}{h_1 + h_\epsilon} \right) \mu_1 + \left(\frac{h_\epsilon}{h_1 + h_\epsilon} \right) \eta, \quad (2)$$

whereas in the pure Bayesian case (no confirmation bias), $E[w_t(y^{t-1})] \xrightarrow[t \rightarrow \infty]{} \eta$. Therefore, there will be a *persistent* effect of the initial belief on wages.

Let w^{LR} denote the (expected) long-run wage (the right-hand side of Equation (2)). The weight of the initial belief μ_1 decreases in the signal-to-prior precision ratio h_ϵ / h_1 . This means that, ceteris paribus, a relatively smaller initial prior precision h_1 (larger prior variance) would move w^{LR} closer to a worker's true productivity. However, the very idea of a significantly large prior variance $1/h_1$ seems to be at odds with the notion of confirmation bias. Increasing the signal precision h_ϵ has the same effect on w^{LR} as reducing h_1 .

We define the retribution bias (RB) as the difference between the expected wage with and without confirmation bias.⁷ It is straightforward to show that

$$RB \xrightarrow[t \rightarrow \infty]{} (w^{LR} - \eta) = \frac{h_1}{h_1 + h_\epsilon} (\mu_1 - \eta) \quad (3)$$

As shown by Equation (3), in the long run, the retribution bias is determined by the prior-mean error (PME) and the entrenchment of beliefs. The PME is measured by the distance between the initial (subjective) assessment of productivity and the true productivity of the worker, $(\mu_1 - \eta)$. If potential employers are confirmation-biased and hold negative initial beliefs about the productivity of the worker ($\mu_1 < \eta$), there will be persistent under-retributions of his or her productivity, no matter how large the number of signals is. The entrenchment of beliefs is captured by the ratio $h_1 / (h_1 + h_\epsilon)$. The higher the entrenchment (the degree of 'conviction' about one's initial assessments), the greater the retribution bias that will persist given a PME.

⁷In the limit, RB can be interpreted as a measure of economic discrimination due to confirmation-biased learning. In the limit, $RB < 0$ implies that the worker is not receiving 'pay or remuneration commensurate with their productivity' ([Aigner & Cain, 1977](#)), as measured by η .

3 Discrimination in the labor market

We now extend our learning model with confirmation bias to study patterns of discrimination. Assume, for simplicity, that workers belong to two different social groups or categories (such as gender, ethnicity, or any other), indexed by $g \in \{M, m\}$. Group M is the majority group, while m makes reference to any minority or disadvantaged group facing a negative stereotype (like women, Blacks, Muslims, immigrants, etc.). Employers cannot observe individual worker productivity, but since group membership is observable they can condition their initial beliefs on g , that is:

$$\eta^g \sim \mathcal{N}(\mu_1^g, 1/h_1^g),$$

where μ_1^g is the prior mean and h_1^g is the prior precision for group $g \in \{M, m\}$. We also allow the precision of the signal to differ across groups, and write h_ε^g . To emphasize the role of group identity in discrimination, we will assume that both groups have the same average true productivity, but that employers have negative stereotypes (generalizations about groups that are applied to individual group members simply because they belong to that group; [Heilman, 2012](#)) about the minority group: minority workers are seen as less productive, so that $\mu_1^m < \mu_1^M$.^{8,9}

Consider two individuals with the same true productivity $\bar{\eta}$, but belonging to different groups. Upon entry to the labor market, there will be (statistical) discrimination against the minority worker, because with no information other than group identity, $w_1^m = \mu_1^m < \mu_1^M = w_1^M$. But will discrimination persist? Most economic models that regard pre-market decisions as exogenous have shown that statistical discrimination can indeed persist, provided the number of signals is limited ([Arrow, 1973](#); [Aigner & Cain, 1977](#); [Altonji & Pierret, 2001](#); [Bohren et al., 2019](#)). However, with Bayesian learning (which those models assume), wages would eventually converge to true productivity as the number of signals grows large, and hence both workers would expect to receive the same wage – discrimination cannot persist in the long run in such a setup. As hinted above, this need not be the case with confirmation-biased learning.¹⁰

Under confirmation-biased learning, the long-run wages for two individuals belonging to different groups but who are otherwise identical (i.e., have equal productivities and output

⁸Our assumption corresponds to what [Bohren et al. \(2019\)](#) call biased belief-based partiality (towards members of group M).

⁹Because groups have the same true average productivity, beliefs are necessarily inaccurate in the sense of [Bohren et al. \(2025\)](#).

¹⁰[Elmslie & Sedo \(1996\)](#) and [Goldsmith et al. \(2004\)](#) propose an alternative mechanism that produces long-run effects of initial discrimination based upon Festinger's theory of cognitive dissonance. In response to discrimination, workers may adjust their beliefs about the quality of the job that they can expect to attain, thereby reducing their labor supply. Early discrimination may then negatively affect future employability and wages. The model focuses on the effects of initial discrimination on the supply side of labor, whereas ours focuses on biased beliefs by employers.

histories) will be:

$$\begin{aligned} w_M^{LR} &= \mu_1^M \left(\frac{h_1^M}{h_1^M + h_\varepsilon^M} \right) + \bar{\eta} \left(\frac{h_\varepsilon^M}{h_1^M + h_\varepsilon^M} \right), \text{ and} \\ w_m^{LR} &= \mu_1^m \left(\frac{h_1^m}{h_1^m + h_\varepsilon^m} \right) + \bar{\eta} \left(\frac{h_\varepsilon^m}{h_1^m + h_\varepsilon^m} \right). \end{aligned}$$

We define long-run discrimination as $LRD \equiv w_m^{LR} - w_M^{LR}$. Therefore,

$$LRD = \left[\mu_1^m \left(\frac{h_1^m}{h_1^m + h_\varepsilon^m} \right) - \mu_1^M \left(\frac{h_1^M}{h_1^M + h_\varepsilon^M} \right) \right] + \bar{\eta} \left[\left(\frac{h_\varepsilon^m}{h_1^m + h_\varepsilon^m} \right) - \left(\frac{h_\varepsilon^M}{h_1^M + h_\varepsilon^M} \right) \right].$$

Letting $x \equiv h_1/h_\varepsilon$ we can write equivalently:

$$LRD = \left[\mu_1^m \left(\frac{x^m}{1 + x^m} \right) - \mu_1^M \left(\frac{x^M}{1 + x^M} \right) \right] + \bar{\eta} \left[\left(\frac{1}{1 + x^m} \right) - \left(\frac{1}{1 + x^M} \right) \right]. \quad (4)$$

A minority worker is discriminated against when her expected long-run wage is lower than the long-run wage of an otherwise identical majority worker; that is, when $LRD < 0$.

Before deriving general conditions for $LRD < 0$, let us analyze particular cases. In the spirit of Phelps (1972), and similar to Coate & Loury (1993) and Bohren et al. (2019), assume that groups only differ in their prior means – i.e., $h_1^M = h_1^m = h_1$ and $h_\varepsilon^M = h_\varepsilon^m = h_\varepsilon$, or $x^m = x^M = x$. Then:

$$LRD = (\mu_1^m - \mu_1^M) \left(\frac{x}{1 + x} \right).$$

It is clear that initial group-level discrimination, $\mu_1^m < \mu_1^M$, leads to persistent between-group discrimination under confirmation bias – i.e., $LRD < 0$ for all x .¹¹ The magnitude of the discrimination will depend on the entrenchment of beliefs, as measured by $x/(1 + x) = h_1/(h_1 + h_\varepsilon)$. A relatively high h_1 (stronger *a priori* convictions about initial beliefs) leads to more discrimination for a given difference in initial beliefs, whereas a higher h_ε (better chances to convey information through less noisy signals) moves LRD towards zero.

An alternative set of parameters explored in the literature assumes, contrary to what we have done here, that prior means are identical across groups $\mu_1^m = \mu_1^M = \mu_1$ (so that individuals are perceived to have the same average productivity), but that precisions differ. In that case,

$$LRD = (\mu_1 - \bar{\eta}) \left(\frac{x^m}{1 + x^m} - \frac{x^M}{1 + x^M} \right).$$

For instance, signals might be noisier for minority workers (Phelps, 1972; Aigner & Cain, 1977; Lundberg & Startz, 1983; Lundberg, 1991; Cornell & Welch, 1996; Farmer & Terrell, 1996), and hence $h_\varepsilon^m < h_\varepsilon^M$. Lower signal precision could also be interpreted as greater subjectivity in judgment, as in Bohren et al. (2019). Alternatively, potential employers might think that minority group members tend to resemble each other quite a bit, and thus $h_1^m > h_1^M$: the

¹¹Discrimination can also arise as the result of favoritism or partiality towards the in-group rather than hostility against out-groups. Ahmed (2007) shows experimental evidence of this.

minority group has a lower perceived variability (Park & Judd, 1990).¹² Either assumption (and the existing evidence) suggests that $x^m > x^M$ is the most relevant case. Then there will be long-run discrimination as long as the PME is negative for these individuals – i.e., whenever $\mu_1 < \bar{\eta}$ and individuals' productivities are under-appreciated by the market at the outset.

We can now turn to the exploration of the general case (with $\mu_1^m < \mu_1^M$). To that end, it is useful to write Equation (4) as follows:

$$\text{LRD} = (\mu_1^m - \bar{\eta}) \frac{x^m}{1 + x^m} - (\mu_1^M - \bar{\eta}) \frac{x^M}{1 + x^M}. \quad (5)$$

Discrimination, a feature of behavior, depends on the primitives of the model, and can go either way. Notice in particular that $\mu_1^m < \mu_1^M$ is not sufficient to have $\text{LRD} < 0$, and that reversals in discrimination are possible.¹³ The following proposition summarizes the results of a sensitivity analysis:

Proposition 1: *Long-run discrimination against the minority worker ($\text{LRD} < 0$):*

- *obtains whenever $\mu_1^m < \bar{\eta} < \mu_1^M$, for any x^m, x^M , and is increasing in both x^m and x^M in this case;*
- *is more likely the lower μ_1^m and the higher μ_1^M ;*
- *is more (less) likely the higher is x^m whenever $\mu_1^m < \bar{\eta}$ ($\mu_1^m > \bar{\eta}$); and*
- *is less (more) likely the higher is x^M whenever $\mu_1^M < \bar{\eta}$ ($\mu_1^M > \bar{\eta}$).*

The proof follows from simple inspection of Equation (5) and is omitted.

Proposition 1 helps us think about the expected effects of policies aimed at reducing discrimination under confirmation bias. Procedures that involve hiding the group identity of the worker at the time of evaluation (like the blind auditions analyzed by Goldin & Rouse, 2000; or the double-blind refereeing in academic journals discussed by Blank, 1991) would be akin to making $\mu_1^m = \mu_1^M$. Such interventions would eliminate discrimination upon entry to the labor market and would reduce discrimination in the long run –but would not eliminate discrimination unless $x^M = x^m$. Empathy training programs have been shown to reduce stereotyping (Galinsky & Moskowitz, 2000), which would amount to bringing μ_1^m closer to μ_1^M in our setup, and we predict that this intervention would reduce discrimination, consistent with evidence (McGregor, 1993; Batson et al., 1997; Stephan & Finlay, 1999; Galinsky & Moskowitz, 2000).

Attracting individuals' attention to within-group differences in the minority group would imply a reduction in μ_1^m and hence in x^m . This approach has the potential to reduce discrimination if the minority worker was under-appreciated to begin with (i.e., if $\mu_1^m < \bar{\eta}$). Evidence

¹²This may be related to the outgroup homogeneity effect (Judd & Park, 1988; Linville, 1998) –the tendency for people to see outgroup members as more alike than ingroup members– if potential employers belong to the majority group.

¹³For example, set $\mu_1^M = 3$, $\mu_1^m = 2.8$, $\bar{\eta} = 1$, $x^M = 1$, $x^m = 2$. Then $\mu_1^m - \mu_1^M = -0.2$ but $\text{LRD} = +0.2$.

from laboratory and field experiments suggests that increasing the perceived variability of the minority group indeed reduces discrimination – it might even reduce stereotyping, which would reinforce the effect (see, e.g., [Brauer & Er-rafy, 2011](#)). Reducing noise in the measurement of minority workers' output – that is, increasing h_ϵ^m – would also imply a reduction in x^m , and have similar effects on discrimination. Evidence consistent with reduced discrimination when minority workers' signals are interpreted with greater precision is provided in [Sarsons et al. \(2021\)](#) and [Bohren et al. \(2019\)](#). Increased h_ϵ^m can also be interpreted as reduced ambiguity, which makes discrimination less likely to arise according to evidence in [Nieva & Gutek \(1980\)](#) and [Heilman & Haynes \(2007\)](#).

The LRD in Equation (4) is an absolute-level wage gap, and it is a function of the workers' underlying productivity, $\bar{\eta}$. As one moves up the income distribution (i.e., as the wages of *both* workers, w_m^{LR} and w_M^{LR} , increase), the relative size of this gap might change. Since a simultaneous increase in both wages can only occur through an increase in the workers' productivity, the question boils down to whether the absolute wage gap, LRD, increases or decreases with underlying productivity, $\bar{\eta}$.¹⁴ Taking the derivative of Equation (4) with respect to $\bar{\eta}$ quickly shows that:

$$\frac{\partial \text{LRD}}{\partial \bar{\eta}} < 0 \iff x^m > x^M \quad (6)$$

As we have discussed previously, $x^m > x^M$ is the most likely case. Hence, as one moves up the income distribution, the wage gap would be reduced. If we associate higher wages and higher productivity with higher skill jobs, then this prediction fits nicely with the existing evidence on the wage gap between blacks and whites in the US –one of the main stylized facts found by [Lang & Lehmann \(2012\)](#) in their extensive review of racial discrimination in the labor market. To the extent that more education is associated to higher productivity, the prediction also squares well with the evidence in [Nopo et al. \(2012\)](#) on international gender wage gaps.

4 Discrimination and segregation

In the previous section we have characterized patterns of wage discrimination in labor markets in which all employers statistically discriminate based on the same (incorrect) beliefs and suffer from confirmation bias. A natural and important next step to consider would be a situation where employers in the same labor market are heterogeneous in terms of their initial beliefs (for example, employers who have the same group identity as a potential employee may have different initial beliefs about worker productivity from employers who do not). Such an extension would also allow us to discuss whether the same set of factors that produce a pay differential would also produce segregation ([Blau & Jusenius, 1976](#)). Pay gaps and occupational segregation by group membership have long been associated with employment discrimination ([Bergmann, 1974](#)) and tend to coexist in the labor market. Furthermore, occupational segregation by gender, for instance, explains most of the gender wage gap ([Bielby & Baron, 1986](#); [Groshen, 1991](#); [Blau](#)

¹⁴We thank an anonymous reviewer for suggesting this comparative statics exercise.

& Kahn, 2017).

We continue to assume that workers belong to group $g \in \{M, m\}$, and that employers cannot observe individual worker productivity, but only their group belonging. Each firm has only one position to fill, and hence, can hire at most one worker in each period. A firm that does not hire a worker makes zero profits in that period. We model the heterogeneity of employers by assuming that they differ in their initial beliefs about each type of worker, μ_{fg} . For simplicity, we assume precisions are equal across employer types $f \in \{M, m\}$.¹⁵

As before, labor supply is perfectly inelastic, so wages are determined solely by the demand side, and we assume workers are interchangeable in production. Employers, on the other hand, hold different initial beliefs on worker productivity, i.e., $\mu_{Mg} \neq \mu_{mg}$ for some g . Let $F_f > 0$ denote the number of type f firms, $F \equiv F_M + F_m$, and $\pi^f \equiv F_f/F$. Equivalently for workers, we define $L \equiv L_M + L_m > 0$, and $p^g \equiv L_g/L$. We assume $F > L$ (there are more firms than workers overall), and $p^m > \pi^m$ (e.g., because minority workers are underrepresented in managerial positions).

We consider a competitive labor market with no search costs, so that in each period all firms can make bids to all workers at no cost. Given our assumptions, workers need only look at current wage offers and accept the best offer in each period. If an outside offer ties the offer of the current employer, we assume the worker stays.¹⁶

Firms compete à la Bertrand in wage offers each period. To determine wage offers by firms, first note that in any period $t > 0$ the willingness to pay (WTP) of a type- f firm for the services of a worker of group g with a signal history y_i^{t-1} is determined by its current confirmation-biased belief about that worker's productivity, which is given by:

$$\hat{\mu}_t(\mu_{fg}, y^{t-1}) = \mu_{fg} \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\epsilon) - h_\epsilon} \right) + \hat{v}_t(y^{t-1}) \quad (7)$$

where

$$\hat{v}_t(y^{t-1}) \equiv \sum_{j=1}^{t-1} y_j \left(\frac{2^{t-1-j} h_\epsilon}{2^{t-1} (h_1 + h_\epsilon) - h_\epsilon} \right)$$

summarizes the *value of a history of signals at t* (i.e., its contribution to the posterior belief). For any t , conditional on a history y^{t-1} , the WTP is strictly increasing in the firm's prior μ_{fg} . Moreover, two workers with the same output history would receive different wage offers to the extent that the firm's prior beliefs on them differ.¹⁷

Bertrand competition implies that wages offered by all firms to any given worker (of group g with history y^{t-1}) will be determined by the WTP of the *marginal* employer of that social group. The marginal employer of group g will be the firm type with the highest WTP for such a worker if there is no ingroup congestion for group g ($L_g < F_g$). If there is congestion in one

¹⁵In choosing the same labels for worker and employer groups we are focusing on employers who may or may not have the same group identity as potential employees. Alternatively, different employers could have the same beliefs within groups, but with different groups interpreted as different occupations (like 'masculine' and 'feminine').

¹⁶Alternatively, we can assume there exists an infinitesimal cost of changing jobs.

¹⁷Once again, due to confirmation bias, differences will persist in the long run.

group, the marginal employer is the firm type with the lowest WTP, and the other firm type can lower its equilibrium wage offer to that of the marginal employer and enjoy a positive expected profit.

Many cases can arise depending on the ordering of the different initial beliefs of employers. For the sake of brevity, we focus here on the cases we regard as most relevant or plausible.¹⁸ To this end, we assume the following:

- (A1) $\mu_{MM} > \mu_{Mm}$ (i.e., employers in the majority group regard majority workers as more productive on average than minority workers)
- (A2) $\mu_{mm} > \mu_{mM}$ (i.e., minority workers are regarded more favorably by minority employers than by majority employers)
- (A3) $\mu_{MM} > \mu_{mM}$ (i.e., majority workers are regarded more favorably by majority employers than by minority employers)

This ordering corresponds to a case in which prior beliefs favor ingroup matching (homophily): for each worker group g , the ingroup employer has a higher prior than the outgroup employer (and, conditional on history, a higher WTP). Other assumptions might well generate different predictions, but ingroup favoritism seems like a natural starting point for exploring the consequences of confirmation bias in a setting with heterogeneous employers—and there is plenty of evidence (see, e.g., [Lewis & Sherman, 2003](#); [Nunley et al., 2011](#); [Doleac & Stein, 2013](#); [Wright & Schepker, 2015](#); [Whysall, 2018](#); [Kline et al., 2022](#)) that such ingroup biases are pervasive in this context, making it a plausible assumption.

Suppose first that there is no ingroup congestion (NIC), i.e., $F_g > L_g \forall g$. Then, under (A1)–(A3) there exists a unique full segregation equilibrium in which all firms offer wages that are equal to their current beliefs $\hat{\mu}_t(\mu_{fg_i}, y_i^{t-1})$ about the productivity of each worker, M-workers accept offers from M-firms (which are higher than those offered to them by m-firms), and m-workers accept offers from m-firms (which are higher than those offered to them by M-firms).¹⁹ In this equilibrium all firms make zero expected profits and there will be discrimination except in the case in which $\mu_{MM} = \mu_{mm}$.

In contrast, if NIC is violated (i.e., there is a worker group g that saturates job slots, $L_g > F_g$), then under (A1)–(A3) there is a partial segregation equilibrium.²⁰ For concreteness, assume it is the minority group that cannot find jobs only in minority firms. In such an equilibrium all M-workers are employed by M-firms, which pay their WTP and hence make no expected profits. Some m-workers cannot be employed by m-firms due to slot constraints, and hence M-firms become their marginal employer. This drives m-firms' wage offers down to that of M-firms, and all firms pay in equilibrium a wage equal to the lowest WTP, that of the M-type.

¹⁸The analysis of every case not covered here is available from the authors upon request.

¹⁹While segregation depends on beliefs in the model, beliefs could in turn be affected by segregation. See [Levy & Razin \(2017\)](#) for a model of the coevolution of segregation, beliefs, and discrimination.

²⁰[Duncan & Duncan \(1955\)](#)'s index of dissimilarity (or segregation) would be equal to $(F_g/L_g) * 100$, which is below 100.

m-workers employed by M-firms receive a lower wage than M-workers with a similar history of signals employed by the same type of firms (because $\mu_{MM} > \mu_{Mm}$ by (A1)). m-workers employed by m-firms are also discriminated against, as they receive wages lower than comparable M-workers in M-firms. Therefore, minority workers are discriminated against by both ingroup and outgroup firms.

When firms have heterogeneous prior beliefs, the equilibrium wage gap between M and m workers with the same history y^{t-1} is given by:

$$\begin{aligned} & w_t^m(m, y_i^{t-1}) - w_t^M(M, y_i^{t-1}) \\ &= \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\varepsilon) - h_\varepsilon} \right) (\mu_{Mm} - \mu_{MM}) \end{aligned} \quad (8)$$

under partial segregation, and by

$$\begin{aligned} & w_t^m(m, y_i^{t-1}) - w_t^M(M, y_i^{t-1}) \\ &= \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\varepsilon) - h_\varepsilon} \right) (\mu_{mm} - \mu_{MM}) \end{aligned} \quad (9)$$

under full segregation. The full segregation wage gap is negative for any t (discrimination against minority workers) if and only if $\mu_{mm} < \mu_{MM}$, whereas it is always negative under partial segregation by (A1). Whatever the equilibrium, discrimination is a function of a) the entrenchment of prior beliefs, and b) the difference in prior means of the *marginal employers* of m- and M-workers.

Long-run discrimination is given by

$$\lim_{t \rightarrow \infty} \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\varepsilon) - h_\varepsilon} \right) (\mu_{mm} - \mu_{MM}) = \left(\frac{h_1}{h_1 + h_\varepsilon} \right) (\mu_{mm} - \mu_{MM}) \quad (10)$$

under full segregation, and by

$$\lim_{t \rightarrow \infty} \left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\varepsilon) - h_\varepsilon} \right) (\mu_{Mm} - \mu_{MM}) = \left(\frac{h_1}{h_1 + h_\varepsilon} \right) (\mu_{Mm} - \mu_{MM}) \quad (11)$$

under partial segregation.

In a cross section of occupations, after controlling for relevant differences across workers, discrimination against minority workers should be more pronounced in partially segregated occupations than fully segregated ones for any t . From (8) and (9), the difference between the short-run partial-segregation and the short-run full-segregation wage gaps is given by

$$\left(\frac{2^{t-1} h_1}{2^{t-1} (h_1 + h_\varepsilon) - h_\varepsilon} \right) (\mu_{Mm} - \mu_{mm}),$$

which is negative by (A2).

The comparison between the full- and partial-segregation equilibria shows that M-firms pay less to their workers on average under partial segregation, which squares well with the substantial evidence showing a negative correlation between the average pay (of men and

women combined) and the proportion of all employees in that job that is female, if we take women as the minority group; see, e.g., [Killingsworth \(1987\)](#), [Macpherson & Hirsch \(1995\)](#), and [Boraas & Rodgers III \(2003\)](#).

In a time series, both in full- and partial-segregation equilibria, discrimination against m-workers is worse in the short run, consistent with evidence showing that labor-market experience mitigates discrimination against, e.g., immigrants ([Baert et al., 2017](#); [Fays et al., 2021](#)).

5 Discrimination and competition

Wherever stereotypes come from (e.g., from employers using the representative heuristic of [Bordalo et al., 2016](#)), our model shows how confirmation-biased employers tend to perpetuate their consequences for wage gaps. Hence, discriminatory behavior survives the passing of time. But will discriminators survive in the long run? This is a related, but different, question. As [Schwartzstein \(2014, p. 1447\)](#) emphasizes, “the logic of confirmation bias does not by itself pin down which incorrect beliefs we can expect to persist”. Thus, something else is needed to pin down beliefs. Competition among employers might achieve precisely that.

The issue has been discussed at least since [Becker \(1957\)](#) in the context of taste-based discrimination, where relative wages of minority workers are determined by the most prejudiced employer with whom they come into contact –the marginal discriminator ([Charles & Guryan, 2008](#)). Therefore, it is the marginal, rather than the average, prejudice of employers that determines the wage gap.

This logic of the marginal competitor applies equally well to models of statistical discrimination like ours (see, e.g., [Nunley et al., 2011](#)). To be concrete, assume n potential employers have only one position to fill, and hence, can hire at most one worker in each period. A firm that does not hire a worker makes zero profits in that period. Employers bid each period for the services of a worker from group $g \in \{M, m\}$. All employers have the same belief about the productivity of M-workers, $\mu_i^M = \mu^M$ for all i , and they only differ in the belief they have on the productivity of m-workers, μ_i^m ($i = 1, \dots, n$). In particular, employers can be ranked according to how biased they are against m-workers:

$$\mu_i^m = \mu^M - b_i$$

with n being the most biased employer, and $b_i \in (0, \mu^M)$ for all i .

Firms compete à la Bertrand in wage offers each period. For any t , conditional on a history y^{t-1} , the willingness to pay (WTP) is strictly increasing in the firm’s prior μ_i^g . Moreover, two workers with the same output history would receive different wage offers to the extent that the firm’s prior beliefs on them differ. Irrespective of actual employer, the wage of a majority worker will be determined by μ^M , but the wage of a minority worker will be determined by the WTP of the second highest bidder –the marginal employer –i.e., $\mu_2^m = \mu^M - b_2$. In general, with n firms bidding for the services of $k < n$ workers of group m , whether offers for workers are all simultaneous or sequential, the worst-off m-worker’s wage will be determined

by $\mu_{k+1}^m = \mu^M - b_{k+1}$.

The marginal employer of group m will then be the $(k + 1)$ -th firm. Increased competition in the labor market, construed as an increase in the number of employers, will reduce discrimination if it reduces the bias of the marginal employer –i.e., if entrants come from the left tail of the employer bias distribution (the most likely case, since they are the ones who can profit from entry). Because less biased employers have a higher willingness to pay, increasing the number of employers in the labor market raises the probability that the k firms with the highest willingness to pay are less biased, and thus raises the wage offer (Nunley et al., 2011).

Employers with more extreme biases are “driven out of the market” as competition increases (they cease to be the marginal employer). We thus expect labor markets with higher competitive pressure on the demand side to show smaller wage gaps (consistent with evidence in, e.g., Meng, 2004, Nunley et al., 2011, Caminade et al., 2012, and Doleac & Stein, 2013).

Our setup is more amenable to thinking about the consequences of competition in the labor market than in the product market in which our employers operate. In the context of taste-based discrimination, prejudiced employers sacrifice profits by discriminating. As the usual argument goes, such employers would be ultimately driven from the market in the long run in a competitive setting (Arrow, 1972; Comanor, 1973). More recent work has shown that wage gaps can persist in the long run if there is some form of imperfect information, imperfect competition, or adjustment costs (Charles & Guryan, 2008). All these arguments refer to competition in the product market. It is not clear, however, what the relationship between product-market competition and statistical discrimination is (Berkovec et al., 1998; Heyman et al., 2013; Cooke et al., 2019; Fays et al., 2021). For instance, if increased competitive pressure forces firms to improve management practices, it could improve the screening of job candidates, reducing the reliance on group stereotypes (Heyman et al., 2013; Cooke et al., 2019). Improved screening could be interpreted as a reduced χ^m in our model. Whether this would reduce (or increase) discrimination, however, depends also on the other parameters of the model, as can be seen in Equation (5).

6 Concluding remarks

We have developed a model of inaccurate statistical discrimination that highlights confirmation bias as a source of discrimination. We have shown that discrimination in the long run depends not only on the stereotypes that apply to different social groups, but also on how potential employers evaluate a worker’s output and how willing they are to modify their initial beliefs. We have used the model to predict how different interventions to reduce discrimination would work in the face of confirmation-biased learning, and to show how the same factors that generate discrimination can also produce segregation.

The limitations of our modeling assumptions are worth mentioning at this point. Our focus has been on labor-market discrimination, which means that we have omitted a worker’s pre-labor market human capital investments (see, e.g., Coate & Loury, 1993; Farmer & Terrell, 1996). Because we had a single job in the analysis, we were not able to address job assignment and promotions, which are important ingredients of a fully-fledged theory of wage dynamics and

careers in organizations (Gibbons & Waldman, 1999a) —extending our model in this direction, for instance along the lines of (Gibbons & Waldman, 1999b), would be straightforward and is an interesting avenue for future research.

We have explored the consequences of incorrect beliefs and confirmation bias for wage discrimination, and generated several predictions. We view this as a theoretical article and have not attempted a rigorous empirical evaluation. However, to flesh out the analysis, we have reported some evidence from other empirical studies that is consistent with those predictions, and may be useful in thinking about the determinants of wage discrimination. It is beyond the scope of our article to provide more systematic evidence on the predictions of the model, but designing an experiment with appropriate parameterizations to directly test the theoretical predictions should be high on the research agenda. Since our model refers to the persistence of wage discrimination, rather than to binary hiring decisions (as in much of the empirical work on non-Bayesian discrimination), we must recognize some of the challenges involved in designing such an experiment, in particular, the elicitation of long-run beliefs.

References

- Ahmed, A. M. (2007). Group identity, social distance and intergroup bias. *Journal of Economic Psychology*, 28(3), 324–337.
- Aigner, D. J. & Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review*, 30(2), 175–187.
- Altonji, J. G. & Blank, R. M. (1999). Race and gender in the labor market. In *Handbook of Labor Economics*, volume 3C (pp. 3143–3259). Elsevier.
- Altonji, J. G. & Pierret, C. R. (2001). Employer Learning and Statistical Discrimination. *The Quarterly Journal of Economics*, 116(1), 313–350.
- Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39(6), 1037–1049.
- Arrow, K. J. (1972). Some mathematical models of race discrimination in the labor market. In A. H. Pascal (Ed.), *Racial Discrimination in Economic Life* (pp. p. 187–203). Lexington: Lexington Books.
- Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton: Princeton University Press.
- Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination? *Economics Letters*, 155, 35–38.
- Barron, K., Dittmann, R., Gehrig, S., & Schweighofer-Kodritsch, S. (2025). Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment. *Management Science*, 71(2), 1600–1622.
- Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self–other merging? *Journal of Personality and Social Psychology*, 73(3), 495–509.
- Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.
- Benjamin, D. J. (2019). Errors in Probabilistic Reasoning and Judgment Biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, volume 2. North-Holland.
- Bergmann, B. R. (1974). Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex. *Eastern Economic Journal*, 1(2), 103–110.
- Berkovec, J. A., Canner, G. B., Gabriel, S. A., & Hannan, T. H. (1998). Discrimination, Competition, and Loan Performance in FHA Mortgage Lending. *The Review of Economics and Statistics*, 80(2), 241–250.

- Bertrand, M. & Duflo, E. (2017). Field Experiments on Discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments* (pp. 309–393). North-Holland.
- Bielby, W. T. & Baron, J. N. (1986). Men and Women at Work: Sex Segregation and Statistical Discrimination. *American Journal of Sociology*, 91(4), 759–799.
- Blank, R. M. (1991). The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *The American Economic Review*, 81(5), 1041–1067.
- Blau, F. D. & Jusenius, C. L. (1976). Economists’ Approaches to Sex Segregation in the Labor Market: An Appraisal. *Signs: Journal of Women in Culture and Society*, 1(3, Part 2), 181–199.
- Blau, F. D. & Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, 55(3), 789–865.
- Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2025). Inaccurate Statistical Discrimination: An Identification Problem. *The Review of Economics and Statistics*, 107(3), 605–620.
- Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, 109(10), 3395–3436.
- Boraas, S. & Rodgers III, W. A. (2003). How does gender play a role in the earnings gap? an update. *Monthly Labor Review*, 126(3), 9–15.
- Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753–1794.
- Brauer, M. & Er-rafiy, A. (2011). Increasing perceived variability reduces prejudice and discrimination. *Journal of Experimental Social Psychology*, 47(5), 871–881.
- Cain, G. G. (1986). Chapter 13 The economic analysis of labor market discrimination: A survey. In *Handbook of Labor Economics*, volume 1 (pp. 693–785). Elsevier.
- Caminade, J., List, J. A., Livingston, J., & Pixel, J. (2012). When the weak become weaker: The effect of market power on third degree price discrimination.
- Campos-Mercade, P. & Mengel, F. (2024). Non-Bayesian Statistical Discrimination. *Management Science*, 70(4), 2549–2567.
- Charles, K. K. & Guryan, J. (2008). Prejudice and Wages: An Empirical Assessment of Becker’s The Economics of Discrimination. *Journal of Political Economy*, 116(5), 773–809.
- Coate, S. & Loury, G. C. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, 83(5), 1220–1240.
- Coffman, K., Collis, M. R., & Kulkarni, L. (2024a). Stereotypes and Belief Updating. *Journal of the European Economic Association*, 22(3), 1011–1054.

- Coffman, K., Flikkema, C. B., & Shurchkov, O. (2021). Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, 129, 329–349.
- Coffman, K., Ugalde Araya, M. P., & Zafar, B. (2024b). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Economic Inquiry*, 62(3), 957–983.
- Comanor, W. S. (1973). Racial Discrimination in American Industry. *Economica*, 40(160), 363–378.
- Cooke, D., Fernandes, A. P., & Ferreira, P. (2019). Product market competition and gender discrimination. *Journal of Economic Behavior & Organization*, 157, 496–522.
- Cornell, B. & Welch, I. (1996). Culture, Information, and Screening Discrimination. *Journal of Political Economy*, 104(3), 542–571.
- Darity, W. A. & Mason, P. L. (1998). Evidence on Discrimination in Employment: Codes of Color, Codes of Gender. *Journal of Economic Perspectives*, 12(2), 63–90.
- Darley, J. M. & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1), 20–33.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
- Doleac, J. L. & Stein, L. C. (2013). The Visible Hand: Race and Online Market Outcomes. *The Economic Journal*, 123(572), F469–F492.
- Duncan, O. D. & Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review*, 20(2), 210–217.
- Elmslie, B. & Sedo, S. (1996). Discrimination, social psychology, and hysteresis in labor markets. *Journal of Economic Psychology*, 17(4), 465–478.
- Erkal, N., Gangadharan, L., & Koh, B. H. (2022). Gender Biases in Performance Evaluation : The Role of Beliefs versus Outcomes.
- Evans, J. S. B. T. (1989). *Bias in Human Reasoning: Causes and Consequences*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.
- Eyting, M. (2025). Why do we Discriminate? The Role of Motivated Reasoning. *SSRN Electronic Journal*.
- Farber, H. S. & Gibbons, R. (1996). Learning and Wage Dynamics. *The Quarterly Journal of Economics*, 111(4), 1007–1047.
- Farmer, A. & Terrell, D. (1996). Discrimination, Bayesian Updating of Employer Beliefs, and Human Capital Accumulation. *Economic Inquiry*, 34(2), 204–219.
- Fays, V., Mahy, B., Rycx, F., & Volral, M. (2021). Wage discrimination based on the country of birth: Do tenure and product market competition matter? *Applied Economics*, 53(13), 1551–1571.

- Freeman, S. (1977). Wage Trends as Performance Displays Productive Potential: A Model and Application to Academic Early Retirement. *The Bell Journal of Economics*, 8(2), 419–443.
- Fryer, Jr, R. G., Harms, P., & Jackson, M. O. (2019). Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization. *Journal of the European Economic Association*, 17(5), 1470–1501.
- Galinsky, A. D. & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, 78(4), 708–724.
- Gibbons, R., Katz, L. F., Lemieux, T., & Parent, D. (2005). Comparative Advantage, Learning, and Sectoral Wage Determination. *Journal of Labor Economics*, 23(4), 681–724.
- Gibbons, R. & Waldman, M. (1999a). Careers in organizations: Theory and evidence. In *Handbook of Labor Economics*, volume 3 (pp. 2373–2437). Elsevier.
- Gibbons, R. & Waldman, M. (1999b). A Theory of Wage and Promotion Dynamics Inside Firms*. *The Quarterly Journal of Economics*, 114(4), 1321–1358.
- Goldin, C. & Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90(4), 715–741.
- Goldsmith, A. H., Sedo, S., Darity, W., & Hamilton, D. (2004). The labor supply consequences of perceptions of employer discrimination during search and on-the-job: Integrating neoclassical theory and cognitive dissonance. *Journal of Economic Psychology*, 25(1), 15–39.
- Griffin, D. & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3), 411–435.
- Groshen, E. L. (1991). The Structure of the Female/Male Wage Differential. *The Journal of Human Resources*, 26(3), 457.
- Harris, M. & Holmstrom, B. (1982). A Theory of Wage Dynamics. *The Review of Economic Studies*, 49(3), 315–333.
- Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, 32, 113–135.
- Heilman, M. E. & Haynes, M. C. (2007). Subjectivity in the Appraisal Process: A Facilitator of Gender Bias in Work Settings. In *Beyond Common Sense* chapter 7, (pp. 127–155). John Wiley & Sons, Ltd.
- Heyman, F., Svaleryd, H., & Vlachos, J. (2013). Competition, Takeovers, and Gender Discrimination. *ILR Review*, 66(2), 409–432.
- Judd, C. M. & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, 54(5), 778–788.

- Kahneman, D., Lovallo, D., & Sibony, O. (2019). A Structured Approach to Strategic Decisions. *MIT Sloan Management Review*, 60(3), 67–73.
- Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, 39(1), 98–114.
- Killingsworth, M. R. (1987). Heterogeneous Preferences, Compensating Wage Differentials, and Comparable Worth. *The Quarterly Journal of Economics*, 102(4), 727–742.
- Klayman, J. (1995). Varieties of Confirmation Bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of Learning and Motivation*, volume 32 (pp. 385–418). Academic Press.
- Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic Discrimination Among Large U.S. Employers. *The Quarterly Journal of Economics*, 137(4), 1963–2036.
- Lang, K. & Lehmann, J.-Y. K. (2012). Racial Discrimination in the Labor Market: Theory and Empirics. *Journal of Economic Literature*, 50(4), 959–1006.
- Lepage, L.-P. (2024). Experience-Based Discrimination. *American Economic Journal: Applied Economics*, 16(4), 288–321.
- Levy, G. & Razin, R. (2017). The Coevolution of Segregation, Polarized Beliefs, and Discrimination: The Case of Private versus State Education. *American Economic Journal: Microeconomics*, 9(4), 141–170.
- Lewis, A. C. & Sherman, S. J. (2003). Hiring you makes me look bad: Social-identity based reversals of the ingroup favoritism effect. *Organizational Behavior and Human Decision Processes*, 90(2), 262–276.
- Linville, P. W. (1998). The heterogeneity of homogeneity. In *Attribution and Social Interaction: The Legacy of Edward E. Jones* (pp. 423–487). Washington, DC, US: American Psychological Association.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109.
- Lundberg, S. J. (1991). The Enforcement of Equal Opportunity Laws Under Imperfect Information: Affirmative Action and Alternatives. *The Quarterly Journal of Economics*, 106(1), 309–326.
- Lundberg, S. J. & Startz, R. (1983). Private Discrimination and Social Intervention in Competitive Labor Market. *The American Economic Review*, 73(3), 340–347.
- Macpherson, D. A. & Hirsch, B. T. (1995). Wages and Gender Composition: Why do Women’s Jobs Pay Less? *Journal of Labor Economics*, 13(3), 426–471.
- McGregor, J. (1993). Effectiveness of Role Playing and Antiracist Teaching in Reducing Student Prejudice. *The Journal of Educational Research*, 86(4), 215–226.

- Meng, X. (2004). Gender earnings gap: The role of firm specific effects. *Labour Economics*, 11(5), 555–573.
- Mullainathan, S. & Shleifer, A. (2005). The Market for News. *American Economic Review*, 95(4), 1031–1053.
- Mullainathan, S. & Washington, E. (2009). Sticking with Your Vote: Cognitive Dissonance and Political Attitudes. *American Economic Journal: Applied Economics*, 1(1), 86–111.
- Müller, D. (2010). On Horns and Halos: Confirmation Bias and Job Rotation.
- Narasimhan, C., He, C., Anderson, E. T., Brenner, L., Desai, P., Kuksov, D., Messinger, P., Moorthy, S., Nunes, J., Rottenstreich, Y., Staelin, R., Wu, G., & Zhang, Z. J. (2005). Incorporating Behavioral Anomalies in Strategic Models. *Marketing Letters*, 16(3), 361–373.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2), 175–220.
- Nieva, V. F. & Gutek, B. A. (1980). Sex Effects on Evaluation. *Academy of Management Review*, 5(2), 267–276.
- Ñopo, H., Daza, N., & Ramos, J. (2012). Gender earning gaps around the world: A study of 64 countries. *International Journal of Manpower*, 33(5), 464–513.
- Nunley, J. M., Owens, M. F., & Howard, R. S. (2011). The effects of information and competition on racial discrimination: Evidence from a field experiment. *Journal of Economic Behavior & Organization*, 80(3), 670–679.
- Onuchic, P. (2023). Recent Contributions to Theories of Discrimination.
- Park, B. & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology*, 59(2), 173–191.
- Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, 62(4), 659–661.
- Pouget, S., Sauvagnat, J., & Villeneuve, S. (2017). A Mind Is a Terrible Thing to Change: Confirmatory Bias in Financial Markets. *The Review of Financial Studies*, 30(6), 2066–2109.
- Rabin, M. & Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, 114(1), 37–82.
- Rodgers, W. M. (2009). *Handbook on the Economics of Discrimination*. Edward Elgar Publishing.
- Sarsons, H., Gërxhani, K., Reuben, E., & Schram, A. (2021). Gender Differences in Recognition for Group Work. *Journal of Political Economy*, 129(1), 101–147.
- Schwartzstein, J. (2014). Selective Attention and Learning. *Journal of the European Economic Association*, 12(6), 1423–1452.

- Stephan, W. G. & Finlay, K. (1999). The Role of Empathy in Improving Intergroup Relations. *Journal of Social Issues*, 55(4), 729–743.
- Uhlmann, E. L. & Cohen, G. L. (2007). “I think it, therefore it’s true”: Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2), 207–223.
- Uhlmann, E. L. & Silberzahn, R. (2014). Conformity under uncertainty: Reliance on gender stereotypes in online hiring decisions. *Behavioral and Brain Sciences*, 37(1), 103–104.
- Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.
- Whysall, Z. (2018). Cognitive Biases in Recruitment, Selection, and Promotion: The Risk of Subconscious Discrimination. In V. Caven & S. Nachmias (Eds.), *Hidden Inequalities in the Workplace: A Guide to the Current Challenges, Issues and Business Solutions* (pp. 215–243). Cham: Springer International Publishing.
- Wright, P. M. & Schepker, D. J. (2015). Why Boards Go Awry in Their Hiring Decisions. *Wall Street Journal*.

APPENDICES

A Appendix for Chapter 1

A.1 Theoretical appendix

A.1.1 Proofs

Proof of Lemma 1 Consider v_i everywhere convex. Agent i prefers to avoid information if $\varphi_i > 0$, i.e.

$$v_i(p b + (1 - p) g) > p v_i(b + \Delta) + (1 - p) v_i(g)$$

Since $v_i(\cdot)$ is strictly increasing, we have $v_i(b + \Delta) > v_i(b)$, hence we must have

$$v_i(p b + (1 - p) g) > p v_i(b) + (1 - p) v(g)$$

but this contradicts convexity. Therefore, the agent never avoids information, i.e., the agent is an Always Getter. ■

Proof of Lemma 2 Suppose v_i everywhere concave. First note that the net incentive to avoid information increases in the share of other group members who avoid information:

$$\begin{aligned} \frac{\partial \varphi_i}{\partial \lambda_{-i}} &= p v'_i(p b + (1 - p) g) \frac{\partial b}{\partial \lambda_{-i}} - p v'_i(b + \Delta) \frac{\partial b}{\partial \lambda_{-i}} \\ &= [v'_i(p b + (1 - p) g) - v'_i(b + \Delta)] p \frac{\partial b}{\partial \lambda_{-i}} \\ &> 0 \end{aligned}$$

by concavity of v_i and assumption 1. It follows that:

- If $\varphi_i(\lambda_{-i} = 0) > 0$ then by monotonicity $\varphi_i(\lambda_{-i}) > 0$ for any λ_{-i} , that is, the agent is an Always Avoider.
- If $\varphi_i(1) < 0$ then by monotonicity $\varphi_i(\lambda_{-i}) < 0$ for any λ_{-i} , that is, the agent is an Always Getter.
- If $\varphi_i(0) < 0$ and $\varphi_i(1) > 0$, then by continuity there exists an interior λ_{-i}^* such that the agent acquires information whenever $\lambda_{-i} < \lambda_{-i}^*$ and avoids it when $\lambda_{-i} > \lambda_{-i}^*$ (strategic complementarity). ■

Proof of Lemma 3 We assume throughout that $b(0) + \Delta < \hat{U}_i$. Suppose that if everyone else avoids information, then at date 1 the expected utility under ignorance is so low that it falls below the reference level: $p b(1) + (1 - p) g < \hat{U}_i$. Then v_i is convex at $p b(1) + (1 - p) g$. By convexity,

$$v_i(p b(1) + (1 - p) g) < p v_i(b(1)) + (1 - p) v_i(g)$$

so even without instrumental value of information Δ , avoiding information is dominated by acquiring it. Since v_i increasing, adding the instrumental value makes avoiding information even less desirable:

$$v_i(p b(1) + (1-p)g) < p v_i(b(1)) + (1-p)v_i(g) < p v_i(b(1) + \Delta) + (1-p)v_i(g)$$

Therefore, $\varphi^i(1) < 0$. When all other agents acquire information, agent i prefers to avoid. Suppose that if everyone else acquires information, then at date 1 the expected utility under ignorance goes above the reference level: $p b(0) + (1-p)g > \hat{U}_i$. Then v_i is concave at $p b(0) + (1-p)g$. By concavity,

$$v_i(p b(0) + (1-p)g) > p v_i(b(0)) + (1-p)v_i(g)$$

For avoidance to be optimal, we need

$$v_i(p b(0) + (1-p)g) > p v_i(b(0) + \Delta) + (1-p)v_i(g)$$

Define the difference in expected utilities $d \equiv v_i(g) - v_i(p b(0) + (1-p)g)$. Then, the avoidance condition can be re-expressed as:

$$\begin{aligned} v_i(g) - d &> p v_i(b(0) + \Delta) + (1-p)v_i(g) \\ \text{i.e. } p[v_i(g) - v_i(b(0) + \Delta)] &> d \end{aligned}$$

When $d \rightarrow 0$, the condition above always holds. Therefore, by continuity, there exists a value $d^* > 0$ such that for $d < d^*$ the agent avoids information. If $d < d^*$, the agent avoids information when all other agents in the group acquire it ($\varphi_i(\lambda_{-i} = 0) > 0$) and acquires information when all others avoid it ($\varphi_i(\lambda_{-i} = 1) < 0$). By continuity of $\varphi_i(\lambda_{-i})$, there exists an interior threshold λ^{**} such that agent i strictly prefers to avoid information if $\lambda_{-i} < \lambda_i^{**}$ and strictly prefers to acquire information if $\lambda_{-i} > \lambda_i^{**}$ (strategic substitutability). ■

A.2 Additional analyses

All subsequent analyses are referenced in the document. Brief descriptions are provided for each analysis.

A.2.1 Descriptive statistics

Descriptive statistics of the experimental sample are reported below.

Table A.1: Sample summary statistics

	N	Mean	SD	Median	Min	Max
<i>Demographic characteristics</i>						
Age	460	22.509	3.713	22.000	18.000	47.000
Gender (Female=1)	465	0.583	0.494	1.000	0.000	1.000
Has work experience	460	0.154	0.362	0.000	0.000	1.000
<i>General information preferences</i>						
IPS (overall scale)	464	3.116	0.414	3.111	1.000	4.000
<i>Social and economic preferences</i>						
Patience	464	-0.000	1.000	0.201	-3.493	1.124
Risk seeking	464	-0.000	1.000	0.296	-2.678	1.570
Positive reciprocity	464	-0.000	0.803	-0.017	-3.911	1.244
Negative reciprocity	464	0.000	0.836	0.117	-1.833	1.946
Altruism	464	0.000	0.836	0.096	-2.163	1.947
Trust	464	-0.000	1.000	-0.178	-1.662	2.048
<i>Field of study</i>						
Medicine	465	0.213	0.410	0.000	0.000	1.000
STEM	465	0.200	0.400	0.000	0.000	1.000
Humanities	465	0.228	0.420	0.000	0.000	1.000
Economics	465	0.108	0.310	0.000	0.000	1.000
Business	465	0.211	0.408	0.000	0.000	1.000
Other fields	465	0.056	0.230	0.000	0.000	1.000

Note: The number of observations is lower for some variables elicited at the end of the session (e.g., age) because a few participants did not respond. For other variables (e.g., the IPS score and social preferences), one observation is missing due to a participant leaving the session early to attend a lecture.

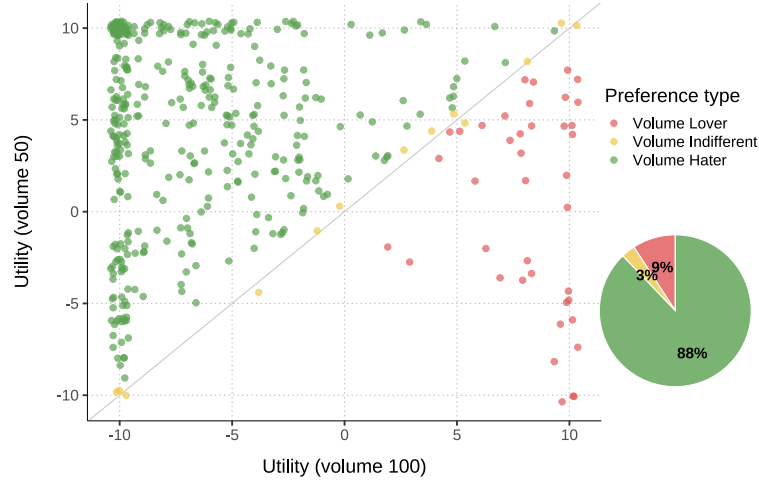
A.2.2 Are screams a bad? Preferences over screams and volume

We classify “Volume Haters” (resp., “Lovers”) as subjects who report higher (lower) utility from hearing screams at volume 50 than at volume 100. Similarly, we classify “Scream Haters” (“Lovers”) as subjects who report a higher (lower) utility from learning that the state is Quiet than that it is Screams. To capture the intensity of these preferences, we compute for each individual the difference between the utility of volume 50 and volume 100, and the difference between the utility of learning that the state is Quiet state relative to Screams.

Figure A.1 shows the joint distribution of subjects' reported utilities of hearing screams at different volumes, distinguishing each type of volume preference. In our sample, 88% of subjects are Volume Haters, 3% are Volume Indifferent, and 9% are Volume Lovers.

Figure A.3 presents the joint distribution of subjects' reported utilities of learning each state, distinguishing each type of scream preference. In our sample, 81% of subjects are Scream Haters, 6% are Scream Indifferent, and 14% are Scream Lovers.

Figure A.1: Preferences over volume of screams: reported utilities and preference types



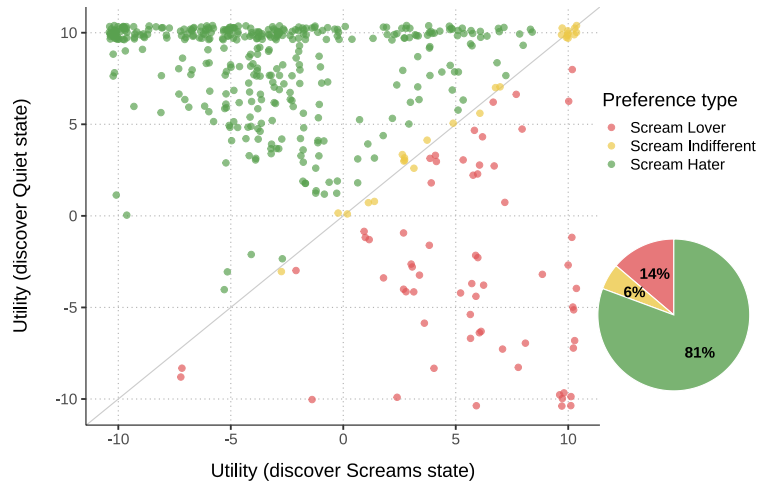
Note: The figure displays reported utilities when screams were played at different volumes. Each point (jittered) corresponds to a participant's reported utilities for volume 50 (y-axis) and volume 100 (x-axis). The dashed 45° line indicates indifference across volumes. Observations are classified into three preference types: Volume Haters (green), who prefer quieter screams; Volume Indifferents (yellow), who report no difference; and Volume Lovers (red), who prefer louder screams. The pie chart summarizes the prevalence of each preference type in the sample.

The joint distribution of preference types over screams and volume is shown in Table A.2.

Table A.2: Distribution of Scream- and Volume-preference types

Scream Preference Type	Volume Preference Type				Total
	Volume Lover	Volume Indif-ferent	Volume Hater		
Scream Lover	23	3	34	4	64
Scream Indifferent	1	5	19	1	26
Scream Hater	16	5	330	24	375
Total	40	13	383	29	465
Scream Lover	4.9%	0.6%	7.3%	0.9%	13.8%
Scream Indifferent	0.2%	1.1%	4.1%	0.2%	5.6%
Scream Hater	3.4%	1.1%	71.0%	5.2%	80.6%
Total	8.6%	2.8%	82.4%	6.2%	100.0%

Figure A.2: Preferences over screams: reported utilities upon learning each state



Note: The figure displays reported expected utilities upon learning the state of the world. Each point (jittered) corresponds to a participant's reported utilities for learning the Quiet state (y-axis) and the Screams state (x-axis). The dashed 45° line indicates indifference across states. Observations are classified into three preference types: Scream Haters (green), who prefer to discover the Quiet state; Scream Indifferents (yellow), who report no difference; and Scream Lovers (red), who prefer to discover the Scream state. The pie chart summarizes the prevalence of each preference type in the sample.

A.2.3 Balance checks

Table A.3 reports the treatment balance checks.

Table A.3: Balance checks

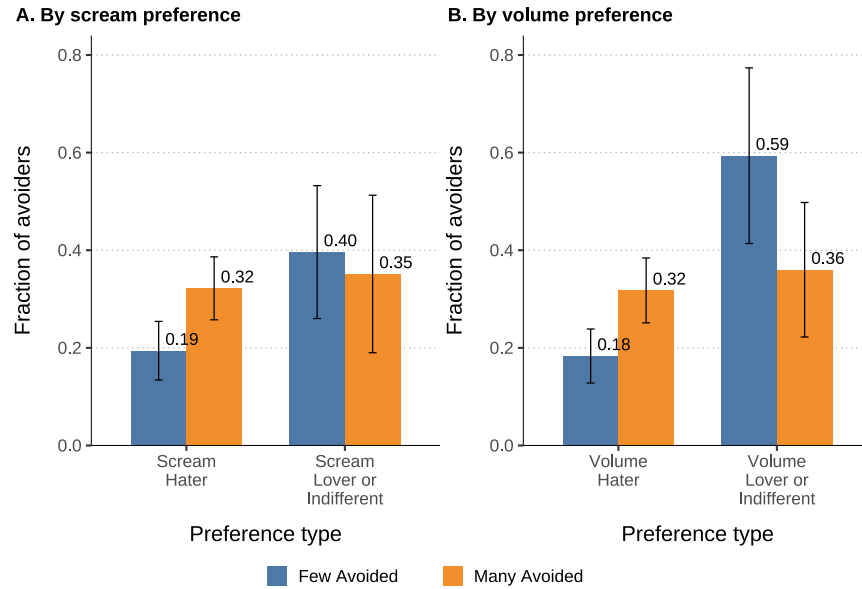
	(1)	(2)	(3)
	Few Avoided	Many Avoided	Difference (2)-(1)
<i>Demographic characteristics</i>			
Age	22.53 (3.98)	22.49 (3.46)	-0.04 (0.35)
Gender (Female=1)	0.57 (0.50)	0.60 (0.49)	0.03 (0.05)
<i>Information preferences</i>			
IPS, overall scale	3.10 (0.42)	3.13 (0.41)	0.02 (0.04)
Health subscale	3.04 (0.72)	3.08 (0.68)	0.04 (0.06)
Finance	2.88 (0.67)	2.90 (0.63)	0.02 (0.06)
Personal	3.11 (0.55)	3.14 (0.55)	0.03 (0.05)
General	3.04 (0.70)	3.08 (0.70)	0.03 (0.06)
Occupational	3.29 (0.48)	3.30 (0.50)	0.01 (0.05)
<i>Field of study</i>			
Medicine	0.19 (0.40)	0.23 (0.42)	0.04 (0.04)
STEM	0.18 (0.39)	0.21 (0.41)	0.03 (0.04)
Humanities	0.19 (0.40)	0.26 (0.44)	0.07* (0.04)
Economics	0.14 (0.35)	0.08 (0.27)	-0.06** (0.03)
Business	0.25 (0.43)	0.18 (0.38)	-0.07* (0.04)
Other fields	0.06 (0.24)	0.05 (0.22)	-0.01 (0.02)
<i>Work experience</i>			
Has work experience	0.16 (0.37)	0.15 (0.36)	-0.01 (0.03)
Observations	223	242	465

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

A.2.4 Robustness to exclusion of scream- and volume-lovers

Here, we show that the main regression results are robust to the exclusion of Scream- and Volume-Lovers. Figure A.3 shows the treatment effect on information avoidance by type of preference for scream (panel A) and for volume (panel B). The treatment effect is positive for scream- and volume-haters, but the direction of the effect is, if anything, the opposite for scream- and volume-lovers. This suggests that, by not excluding scream- and volume-lovers from our main regressions, the treatment effect is rather attenuated.

Figure A.3: Treatment effect on information avoidance by preference over screams and volume



Note: The figure shows treatment effects on information avoidance conditional on individual preference types. Panel A splits participants by scream preference: Scream Haters versus Scream Lovers or Indifferents. Panel B splits by volume preference: Volume Haters versus Volume Lovers or Indifferents. For each subgroup, the bars report the fraction of avoiders under the *Few Avoided* (blue) and *Many Avoided* (orange) treatments, with numerical values displayed above the bars. Error bars represent 95% confidence intervals.

Table A.4 shows our main regression results if we exclude Scream-Lovers. Table A.5 shows our main regression results if we exclude Volume-Lovers. In both subsamples, the magnitude of the estimated treatment effect increases, compared to the full sample (reported in Table 1.2), for all the reported specifications. The treatment coefficient becomes significant at 1% for all specifications.

Table A.4: Treatment effect on information avoidance (subsample of scream haters)

	Dependent variable: Avoided information				
	(1)	(2)	(3)	(4)	(5)
Many Avoided	0.128*** (0.045)	0.162*** (0.052)	0.157*** (0.052)	0.169*** (0.052)	0.170*** (0.053)
Information Preferences Scale			-0.126** (0.056)	-0.124** (0.059)	-0.114* (0.060)
<i>Social preferences</i>					
Altruism				-0.049 (0.031)	-0.072 (0.047)
× Many Avoided					0.050 (0.064)
Positive reciprocity				-0.030 (0.032)	-0.061 (0.046)
× Many Avoided					0.064 (0.066)
Negative reciprocity				0.031 (0.028)	0.037 (0.037)
× Many Avoided					-0.008 (0.054)
Constant	0.194*** (0.031)	-0.041 (0.240)	0.188 (0.324)	0.123 (0.340)	0.112 (0.352)
Strata FEs		✓	✓	✓	✓
Field of study FEs		✓	✓	✓	✓
Individual characteristics			✓	✓	✓
Observations	375	375	371	371	371

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports estimates from a linear probability model regressing information avoidance on the treatment dummy, restricted to the subsample of Scream Haters. See notes in Table 1.2.

Table A.5: Treatment effect on information avoidance (subsample of volume haters)

	Dependent variable: Avoided information				
	(1)	(2)	(3)	(4)	(5)
Many Avoided	0.134*** (0.044)	0.159*** (0.050)	0.161*** (0.049)	0.166*** (0.050)	0.164*** (0.051)
Information Preferences Scale			-0.133** (0.054)	-0.132** (0.057)	-0.127** (0.058)
<i>Social preferences</i>					
Altruism				-0.062** (0.030)	-0.076* (0.045)
× Many Avoided					0.033 (0.061)
Positive reciprocity				-0.023 (0.030)	-0.037 (0.041)
× Many Avoided					0.032 (0.063)
Negative reciprocity				0.021 (0.027)	0.007 (0.035)
× Many Avoided					0.024 (0.054)
Constant	0.183*** (0.028)	-0.136 (0.162)	0.015 (0.270)	-0.088 (0.284)	-0.094 (0.297)
Strata FEs		✓	✓	✓	✓
Field of study FEs		✓	✓	✓	✓
Individual characteristics			✓	✓	✓
Observations	383	383	380	380	380

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports estimates from a linear probability model regressing information avoidance on the treatment dummy, restricted to the subsample of Volume Haters. See notes in Table 1.2

A.2.5 Robustness to alternative specifications

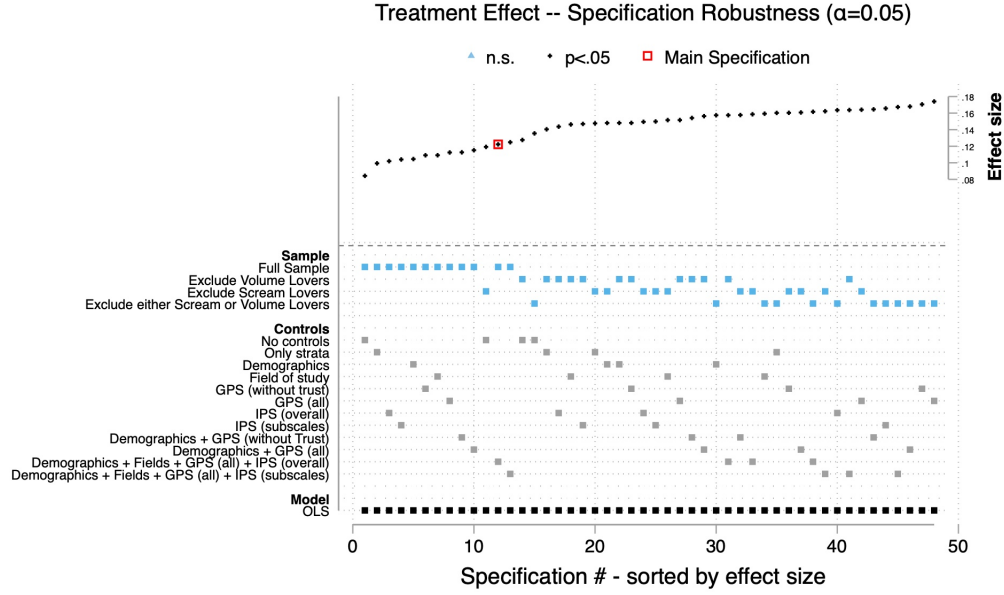
Subsets of controls selected by researcher To assess the robustness of our main treatment effect, we conducted a specification robustness analysis by systematically varying sample restrictions, control sets, and model specifications. Specifically, we estimated the treatment effect across all combinations of four subsamples (full sample, excluding Scream Lovers, excluding Volume Lovers, excluding either) and twelve covariate sets (ranging from no controls to diverse combinations of demographics, field-of-study fixed effects, GPS and IPS measures). The results are presented in Figure A.4. Each point in the top panel represents the estimated effect size under an alternative specification, sorted by magnitude. Black dots indicate that the estimate is statistically significant at 5% level. The lower panels display the corresponding sample, control, and model choices. The main specification is highlighted with a red square. Figure A.5 show a similar plot but at level $\alpha = 0.01$.

Across specifications, the estimated average treatment effect remains consistently positive and is statistically significant in the vast majority of cases. Excluding subgroups (e.g., Volume Lovers, Scream Lovers) or adding or removing controls (demographics, IPS, GPS scales) does not qualitatively alter the result. These findings indicate that the estimated treatment effect is not an artifact of a single specification, but instead remains robust to a broad set of alternatives.

Post-double selection The specification robustness analyses above rely on a combinatorial approach, where pre-defined subsets of covariates are either included or excluded jointly as blocks. This design highlights robustness across broad sets of controls but lacks some flexibility: for instance, social preferences may be either included or excluded as a block, without considering combinations where altruism and reciprocity enter separately. To allow for such flexibility, we turn to the post-double selection procedure (Belloni et al., 2014), a machine-learning-based, data-driven method for selecting controls from a potentially large set of covariates with the goal of causal estimation of parameters.

Results are reported in Table A.6. We estimate the treatment effect on information avoidance using both the plug-in method (columns 1–4) and cross-validation (columns 5–8). All specifications include strata fixed effects, and we progressively add four sets of controls: (i) baseline controls (economic and social preferences; IPS scales, subscales, and individual questions; field of study dummies; age; gender; and work experience), (ii) experimental context and comprehension controls (group size, perceived ease of instructions, and quiz difficulty), (iii) quiz performance until the information decision (questions answered and number correct), and (iv) quiz performance at the end of the experimental session (questions answered and number correct, performance level, and earnings).

Across all specifications, the estimated treatment effect remains positive and statistically significant, with magnitudes similar to those in the main regressions. These findings confirm that the main result (higher information avoidance in the *Many Avoided* condition) is robust to alternative specifications where selection of controls is data-driven.

Figure A.4: Specification robustness plot ($\alpha = 0.05$)

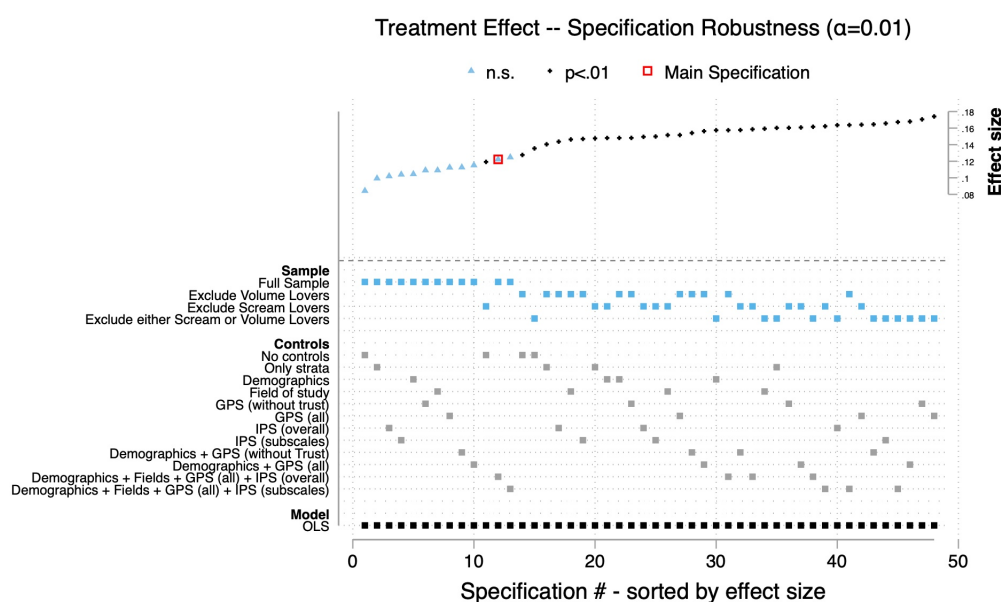
Note: The figure reports a specification curve analysis of the treatment effect. Each point represents the estimated effect under an alternative model specification, sorted by effect size. The top panel plots estimated effect sizes, with black dots indicating specifications significant at the 5% level, blue triangles indicating non-significant estimates, and the red square denoting the main specification reported in Section 5.2. The lower panels display the sample restrictions, control sets, and model choices corresponding to each specification. Effect sizes remain consistently positive across specifications, indicating robustness of the main result.

Table A.6: Post-double selection estimates

	Method: Plug-in				Cross-Validation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Many Avoided	0.095* (0.050)	0.101* (0.056)	0.101* (0.056)	0.101* (0.056)	0.123** (0.049)	0.109* (0.056)	0.110* (0.056)	0.110* (0.057)
Strata FEs	✓	✓	✓	✓	✓	✓	✓	✓
Set 1: Baseline controls	✓	✓	✓	✓	✓	✓	✓	✓
Set 2: Context & comprehension		✓	✓	✓		✓	✓	✓
Set 3: Quiz performance A			✓	✓			✓	✓
Set 3: Quiz performance B				✓				✓

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports post-double selection estimates of the effect of being in the Many Avoided condition on information avoidance. Columns (1)–(4) present results using the plug-in method, while columns (5)–(8) use cross-validation. All specifications include strata fixed effects and progressively add four sets of controls to the covariate selection algorithm: (i) baseline controls (age, gender, field of study dummies, social and economic preferences, Information Preferences Scale (IPS), and work experience), (ii) context and comprehension controls (group size, perceived ease of instructions, and quiz difficulty), (iii) quiz performance before making information decision (number of questions answered and number correct), and (iv) quiz performance at the end of experiment (number of questions answered and number correct, performance level, and earnings).

Figure A.5: Specification robustness plot ($\alpha = 0.01$)

Note: The figure reports a specification curve analysis of the treatment effect. Each point represents the estimated effect under an alternative model specification, sorted by effect size. The top panel plots estimated effect sizes, with black dots indicating specifications significant at the 1% level, blue triangles indicating non-significant estimates, and the red square denoting the main specification reported in Section 5.2. The lower panels display the sample restrictions, control sets, and model choices corresponding to each specification. Effect sizes remain consistently positive across specifications, indicating robustness of the main result.

A.2.6 Alternative outcome: continuous measure of avoidance (IA Scale)

As a robustness check, we replicate our main analysis using the continuous measure of information avoidance (the IA Scale) rather than the binary indicator. Recall that the IA Scale captures the strength of preference for the chosen information decision, ranging from -10 (strong preference for acquiring information) to $+10$ (strong preference for avoiding).

Table A.7 reports the results. The overall pattern closely mirrors that of the main regressions with the binary outcome (Table 1.2): signs of coefficients and statistical significance remain unchanged. Across specifications, the treatment effect of *Many Avoided* remains positive and significant. The coefficient is stable with the inclusion of controls and fixed effects, indicating that our main finding —that information avoidance increases under *Many Avoided*— is robust to the choice of outcome measure.

Moreover, the IA Scale results echo the role of social preferences documented earlier: altruism significantly predicts lower avoidance, while reciprocity remains directionally similar as before, but without though reaching statistical significance. The IPS remains negatively and significantly associated with strength of preference for avoiding information.

Table A.7: Treatment effect on IA Scale

	Dependent variable: IA Scale				
	(1)	(2)	(3)	(4)	(5)
Many Avoided	1.271** (0.624)	1.610** (0.746)	1.677** (0.740)	1.846** (0.743)	1.838** (0.753)
Information Preferences Scale			-1.441* (0.772)	-1.487* (0.807)	-1.389* (0.834)
<i>Social preferences</i>					
Altruism				-1.094** (0.428)	-1.168* (0.638)
× Many Avoided					0.291 (0.873)
Positive reciprocity				-0.326 (0.444)	-0.855 (0.620)
× Many Avoided					1.098 (0.937)
Negative reciprocity				0.224 (0.405)	0.120 (0.586)
× Many Avoided					0.181 (0.820)
Constant	-3.821*** (0.431)	-4.525 (3.463)	-2.052 (4.706)	-2.815 (4.889)	-2.761 (5.065)
Strata FEs		✓	✓	✓	✓
Field of study FEs		✓	✓	✓	✓
Individual characteristics			✓	✓	✓
Observations	465	465	460	460	460

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports estimates from a linear regression of the IA Scale on the treatment dummy. The IA Scale captures the strength of preference for the chosen information decision, ranging from -10 (strong preference for acquiring information) to +10 (strong preference for avoiding). The treatment dummy equals one if the subject was in the *Many Avoided* condition and zero in the *Few Avoided* condition. Regressions include strata fixed effects, demographic controls, field of study, and measures of economic and information preferences (see notes in Table 1.1). The number of observations drops in column (3) to (5) since five participants did not provide information about their age.

A.2.7 Instrumental variable regressions (2SLS)

Table A.8 reports instrumental variable regressions (2SLS) of information avoidance on beliefs about the share of avoiders. Beliefs are instrumented with the treatment assignment. The model specifications (columns 1-5) match those of the main regressions (Table 1.2), except that the treatment dummy is replaced by the instrumented belief.

Once again, the overall pattern is very similar to that of the main regressions (Table 1.2). Across specifications (columns 1–5), a higher belief that others avoid information significantly increases the likelihood of avoiding information. The estimated coefficient on beliefs is stable, positive, and statistically significant across specifications (ranging from $\hat{\beta}^{2SLS} = 0.384$ to 0.604).

The coefficients associated with social preference measures remain directionally similar. Altruism is associated with lower avoidance, while positive reciprocity reduces avoidance in the Few Avoided condition but reverses under the Many Avoided condition, although coefficients lose statistical significance relative to the main regressions.

Finally, the IPS is negatively associated with the decision to avoid informatoin.

Table A.8: Effect of beliefs on information avoidance

	Dependent variable: Avoided information				
	(1)	(2)	(3)	(4)	(5)
Belief about share of avoiders	0.384** (0.181)	0.504** (0.213)	0.523** (0.206)	0.556*** (0.201)	0.604** (0.239)
Information Preferences Scale			-0.123*** (0.045)	-0.122*** (0.046)	-0.065 (0.060)
<i>Social preferences</i>					
Altruism				-0.072*** (0.025)	-0.076 (0.144)
× Belief					0.030 (0.314)
Pos. reciprocity				-0.032 (0.028)	-0.296* (0.165)
× Belief					0.635 (0.386)
Neg. reciprocity				-0.000 (0.024)	-0.080 (0.124)
× Belief					0.091 (0.275)
Constant	0.100 (0.088)	-0.130 (0.232)	0.117 (0.282)	0.006 (0.288)	-0.209 (0.359)
Strata FEs		✓	✓	✓	✓
Field of study FEs		✓	✓	✓	✓
Individual characteristics			✓	✓	✓
Observations	465	465	460	460	460
First Stage F-stat	104.71	75.93	76.49	81.34	12.77

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports estimates from instrumental variable regressions (2SLS) of information avoidance (binary indicator equal to one if the subject avoided information and zero otherwise) on beliefs about the share of avoiders. Beliefs are instrumented using the treatment assignment: the instrument equals one if the subject was in the *Many Avoided* condition and zero if in the *Few Avoided* condition. Regressions include strata fixed effects, demographic controls, field of study, and measures of economic and information preferences (see notes in Table 1.1). The number of observations drops in column (3) to (5) since five participants did not provide information about their age.

A.2.8 Mechanisms — supplementary tables

Table A.9: Interaction of beliefs and information avoidance

	Dependent variable: Avoided information				
	(1)	(2)	(3)	(4)	(5)
Belief about share of avoiders	0.090 (0.173)	0.122 (0.175)	0.098 (0.176)	0.085 (0.173)	0.063 (0.169)
Pr(screams)	-0.194 (0.150)	-0.176 (0.148)	-0.151 (0.148)	-0.156 (0.144)	-0.208 (0.145)
Belief about share of avoiders $\times$ Pr(screams)	0.852*** (0.294)	0.800*** (0.294)	0.787*** (0.297)	0.826*** (0.293)	0.909*** (0.292)
Constant	0.111 (0.076)	0.136 (0.166)	-0.116 (0.215)	0.125 (0.279)	0.029 (0.285)
Strata FEs		✓	✓	✓	✓
Field of study FEs			✓	✓	✓
Individual characteristics				✓	✓
Information preferences				✓	✓
Social and economic preferences					✓
Observations	465	465	465	460	460

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The table reports estimates from a linear probability model regressing information avoidance on the subject's belief about the share of information avoiders among others (i.e., who would choose Later), the subjective belief that the state is Screams ("Pr(screams)"), and their interaction. The regressions include a variety of controls; see notes in Table 1.2.

A.3 Instructions and other screens

Preliminary instructions

At the start of the session, all participants received a printed copy of the instructions below, which were also read aloud by the experimenter.

Welcome to the CeDEx Lab!

Thank you very much for participating in this scientific experiment.

This is an experiment on decision-making. By carefully reading the instructions that will be shown on your computer, you can —depending on the decisions you make— earn an appreciable amount of money. It is therefore important that you pay close attention to the instructions.

We are interested in your individual choices. Therefore, communication is not allowed during the experiment. If you have a question, please just raise your hand.

In addition, there are some ground rules that need to be complied with to ensure fair conditions for all participants:

- **Focus on your own screen.** You should neither interfere with other participants nor look at their workstations (including their screens). Please do not lean backwards and always keep your chair close to your desk.
- **Be patient.** There will be some periods in the experiment in which you will be asked to wait for some time (for example, you might finish your tasks earlier than other people; or you might be asked to just plainly wait). In any case, please remain in your workstation, waiting in silence, without disturbing other participants.
- **No phones.** Please, put your phone(s) in silence mode and put them away (for instance, put them in your coat or bag): do not leave them on your desk. Please do it now. During this session, you should not check or use your phone at any time. If you check or use your phone, you will be asked to leave the session. This is to ensure fair play conditions for all participants.
- **Headphones at all times.** In this experiment, you are asked to wear the headphones provided next to your computer at all times. This is because sounds may be played to you as we will explain later. If you don't properly wear your headphones, you will be asked to leave the experiment.

If you do not comply with any of these ground rules, you will be asked to leave the experiment. Throughout the experiment, some of us, experimenters, will be walking around invigilating for these conditions to be met.

You will soon be given instructions about today's experiment. All instructions will be provided on your computer screen.

Note that participation in this study is entirely voluntary. You can leave the experiment at any time you want. If you leave, we will take into account your progress up to that point to calculate your payment and will pay you accordingly.

In case you have any questions, feel free to ask the experimenter at any time, in a very quiet voice so that no one else is disturbed.

Now, please put on your headphones and then we will be ready to start the experiment.

After these instructions were read aloud, subsequent instructions were presented on participants' computer stations, as described in the next subsection.

Main instructions and decision screens

The following text reproduces the instructions displayed to participants via the computer stations. Each new screen in the original instructions is marked below by a horizontal line. Clarifying notes are in square brackets in italics.

General instructions for participants

Welcome and thank you for participating in this scientific experiment.

First, you will be asked for some contact information. Then, you will be given the instructions on how to earn money in this experiment.

Contact Information

In the following, we will ask for the email address associated with your PayPal account, and your university email address.

This information will be stored only for payment and research purposes.

Please, **be sure that your email addresses are typed correctly. Otherwise, we may not be able to make the payment.**

Please input the email address associated to your PayPal account. [Open text box]

Please type it again. [Open text box]

Now input your university email address. [Open text box]

Input it again. [Open text box]

Please input your **seat number** (this is the number written in the plastic card that you have drawn when you entered into the lab).

[Open text box]

Your consent to participate

Please be informed that the present study involves the possibility of listening to sounds that have been rated as aversive (i.e., unpleasant or disturbing) by participants in other studies. The safety of the participants is a priority for the CeDEx research centre. Therefore, technical measures have been put in place to ensure the hearing safety of the participants.

In order to participate in today's study, your explicit consent to participate is needed. Please note that participation in this study is entirely voluntary.

If you do not give your explicit consent, you will withdraw from this study. There will be no further consequences, and we will keep inviting you to other studies. We will give you £3 as a token of appreciation for coming today.

- ☐ "I do NOT GIVE my explicit consent to participate"
 - ☐ "I GIVE my explicit consent to participate"
-

In this experiment, you and other participants in the room will be part of the same group.

Groups are typically composed of around 30 participants.

Introduction

In today's experiment, you will be asked to do some tasks, which will be explained as you progress through the session. Depending on your performance (and your performance alone) in those tasks, your earnings may increase.

For your participation in today's session, you will receive £5, plus additional earnings based on your performance in the tasks.

Instructions: The Quiz (1/2)

In this experiment, you will be participating in a Quiz. The Quiz questions cover different topics (for instance, sports, geography, history, arts, music, etc.).

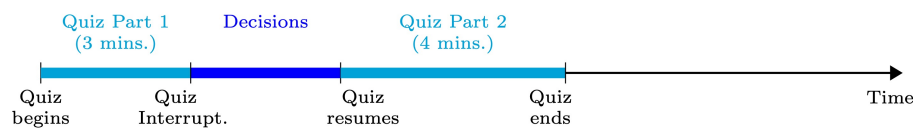
For each question, you will be provided with 4 possible answers, out of which exactly one will be correct.

In this Quiz task, **the more questions you answer correctly, the more you earn.**

The Quiz consists of **two parts**, namely Part 1 and Part 2.

- **Quiz Part 1:** You start with Part 1 of the quiz. You have 3 minutes to answer as many quiz questions as possible.
- **Quiz Interruption and Decisions:** After those 3 minutes, the Quiz will be briefly interrupted. During this interruption, you will be asked to make some decisions, as will be explained on the following screens.
- **Quiz Part 2:** After the interruption, you will answer Part 2 of the quiz, for 4 minutes.

The following figure illustrates the timeline:



The Quiz (2/2)

Below we explain how your earnings are calculated. The key feature is that **you should try to maximise the number of questions answered correctly.**

The number of correct answers in Parts 1 and 2 will be added up to calculate your earnings.

Your earnings increase in accordance with the level you reach. You start at Level 1. To reach Level 2, you need to answer 10 questions correctly (you do not need to correctly answer 10 questions in a row, what counts is the total number of correct answers). Then, to reach each additional level you need to answer correctly 20 extra questions per level. Thus, your earnings from the quiz are determined as follows:

- Level 1 (0-9 correct answers) = £5
- Level 2 (10-29 correct answers) = £6
- Level 3 (30-49 correct answers) = £7
- Level 4 (50-69 correct answers) = £9
- Level 5 (70-89 correct answers) = £13
- Level 6 (90+ correct answers) = £20

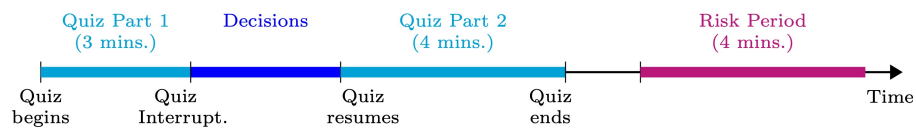
Your earnings from the quiz will be added to the £5 you will earn just for participating in this experiment.

Example: suppose you answer correctly 25 questions in Part 1 and 35 in Part 2. Then the total number of correct answers is 60 ($=25+35$), so the level you reach is 4. Therefore, your earnings from the quiz would be £9. Adding up the £5 from your participation in this experiment, your total earnings would be £14 ($=£9 + £5$). You can see that your earnings increase substantially, the higher the level that you reach.

The Risk Period

After the quiz finishes, a period called the “Risk Period” will start. The Risk Period lasts for 4 minutes.

During the Risk Period, a series of events may or may not happen as explained on the next screens.



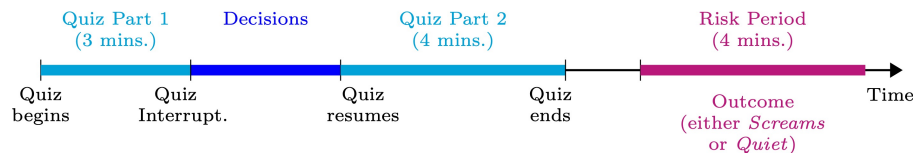
The Outcome

In the Risk Period, one of two outcomes will happen:

- **Screams** outcome: During the Risk Period, your group (including you) will hear a series of **unpleasant distress screams** from your headphones. These will happen at *unpredictable* moments during the Risk Period.

Previous research has shown that people consider similar screams as more aversive (disturbing) than the sound of nails sliding on a chalkboard.

- **Quiet** outcome: you and your group will not hear any screams during the whole experiment.



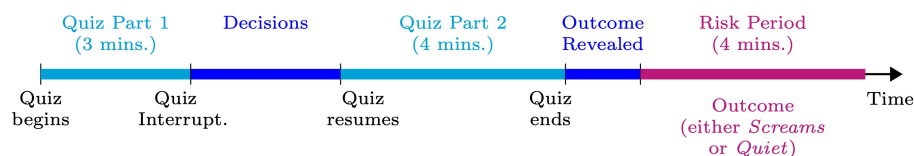
The Lottery

The Outcome (*Screams* or *Quiet*) in the Risk Period is the **same for everyone** in your group, and will be randomly determined by a computerised lottery.

Revelation of Outcome

The Outcome (i.e., *Screams* or *Quiet*) will be **revealed to everyone**, just a **few screens before the Risk Period**.

Therefore, everybody will already know the Outcome before the Risk Period starts.



Your Decision (1/3)

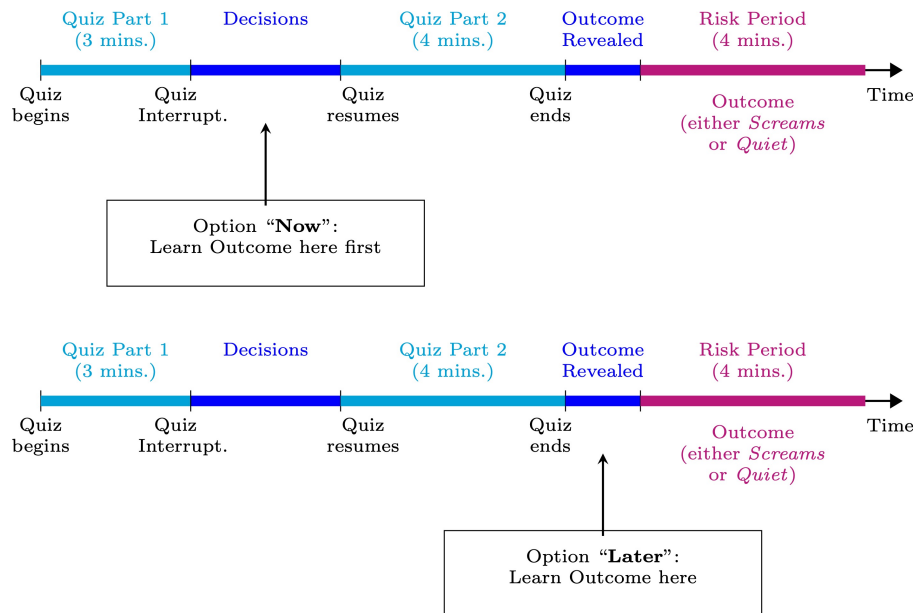
However, you can choose whether you want to know about the Outcome (*Screams* or *Quiet*) earlier or whether you prefer not to be aware of the Outcome until the beginning of the Risk Period.

At the end of Part 1 of the quiz, you will choose one of two options:

- Option “**Now**” implies that you will learn the Outcome (*Screams* or *Quiet*) **immediately** (i.e., during the interruption of the quiz).

- Option “**Later**” implies that you will **not** learn about the Outcome (Screams or Quiet) immediately, and rather you will learn it just before the Risk Period. This means that, during Part 2 of the quiz, you will not know whether in the forthcoming Risk Period you will hear the screams or not.

So the difference between “Now” and “Later” is **whether in Part 2 of the Quiz you are aware** of the Outcome or not. However, the Outcome is not affected by the option you choose.



The Volume of Screams

Although the Outcome is not affected by the option you choose, the **volume** of the screams (in case the Outcome is *Screams*) is affected by the choices that you and everyone else make:

- Each group member who chooses “**Later**” **increases the volume level by about 2 points out of 100.**
- Hence, if everybody chooses “Later”, then the screams will be played at maximum volume (we call this level of volume “level 100”).
- If nobody chooses “Later”, then the screams will be played at volume “level 50”.

This means that, if the Outcome is *Screams*, your choice between “Now” and “Later” would generate a difference in volume of about **2 points, out of 100.**

If instead the Outcome is *Quiet*, your choice has no implications beyond the timing in which you will find out the Outcome.

Check your Understanding

To check your understanding of the instructions, please imagine you are in each of the situations below and answer the following questions.

Situation 1a) Suppose that the Outcome is ***Screams***. In addition, suppose that **everybody** in your group (including you) chooses “**Later**”. What would be the volume of the screams?

- ☐ Level 50
- ☐ Level 52
- ☐ Level 98
- ☐ Level 100
- ☐ There would be no screams

Situation 1b) Suppose that the Outcome is ***Screams***. In addition, suppose that **everybody else** in your group chooses “**Later**”, but **you** choose “**Now**”. What would roughly be the volume of the screams?

- ☐ Level 50
- ☐ Level 52
- ☐ Level 98
- ☐ Level 100
- ☐ There would be no screams

Situation 2a) Suppose that the Outcome is ***Screams***. In addition, suppose that **everybody** in your group (including you) chooses “**Now**”. What would be the volume of the screams?

- ☐ Level 50
- ☐ Level 52
- ☐ Level 98
- ☐ Level 100
- ☐ There would be no screams

Situation 2b) Suppose that the Outcome is ***Screams***. In addition, suppose that **everybody else** in your group chooses “**Now**”, but **you** choose “**Later**”. What would roughly be the volume of the screams?

- ☐ Level 50
- ☐ Level 52

- ☐ Level 98
- ☐ Level 100
- ☐ There would be no screams

Situation 3) Suppose that the Outcome is ***Screams***. In that case:

- Your decision between “Now” and “Later” will...
 - ☐ Only affect the volume of the screams to be heard by you
 - ☐ Only affect the volume of the screams to be heard by participants other than you
 - ☐ Affect the volume of the screams to be heard by everyone (including you)
- The decisions between “Now” and “Later” **by participants other than you** will...
 - ☐ Only affect the volume of the screams to be heard by you
 - ☐ Only affect the volume of the screams to be heard by participants other than you
 - ☐ Affect the volume of the screams to be heard by everyone (including you)

Situation 4) Suppose that the Outcome is ***Quiet***. In addition, suppose that **everybody** in your group (including you) chooses “**Later**”. What would roughly be the volume of the screams?

- ☐ Level 50
- ☐ Level 52
- ☐ Level 98
- ☐ Level 100
- ☐ There would be no screams at all

When you are ready, press the "Continue" button below.

(In case you are not taken to the next screen: some of your answers to the questions above might be wrong, please review them).

Instructions: Your Decision (2/3)

As mentioned, the difference between “Now” and “Later” is whether in Part 2 of the Quiz you are aware of the Outcome or not, and whether the volume of the screams (if the Outcome is Screams) changes by 2 points out of 100.

However, your choice between “Now” and “Later” does **not** change any of the following:

- the duration of the session;
- the sequence of the quiz;
- the amount of time you can spend on the quiz;
- when or whether you may potentially hear the screams.

Finally, note that for this choice between “Now” and “Later”, there is no right or wrong choice. You should simply decide based on your own preferences.

Instructions: Your Decision (3/3)

Recall that if you choose “Now”, then you will be shown a screen that informs you about the Outcome.

Upon being informed about the Outcome, you will be asked to select a button that correctly reflects it. You will need to do it to proceed to the next screens, which include Part 2 of the Quiz.

Differently, if you choose “Later”, then you will be shown a screen that will ask you to wait, and you will not learn the Outcome until Quiz Part 2 is finished, just before the Risk Period.

Control Questions

To briefly check that you understood the instructions correctly, please answer the following questions.

- If you choose option “Now”...
 - ☐ You will learn the Outcome immediately
 - ☐ You will not learn the Outcome immediately, but rather just before the Risk Period
 - ☐ You will never learn the Outcome
- If you choose option “Later”...
 - ☐ You will learn the Outcome immediately
 - ☐ You will not learn the Outcome immediately, but rather just before the Risk Period
 - ☐ You will never learn the Outcome
- If the Outcome is *Screams*, the **timing** in which you will hear the screams...

- will be earlier if you choose option “Now”
- will be earlier if you choose option “Later”
- will be the same independently of whether you choose “Now” or “Later”
- If the Outcome is *Screams*, the **volume** of the screams...
 - will be 2 points higher if you choose option “Later”
 - will be 2 points higher if you choose option “Now”
 - will be the same independently of whether you choose “Now” or “Later”
- Just before the Risk Period starts, you will be informed about the Outcome...
 - Only if you have chosen “Now”
 - Only if you have chosen “Later”
 - In any case, independently of whether you have chosen “Now” or “Later”
 - You will not be informed about the Outcome at that point

When you are ready, press the “Continue” button below.

(In case you are not taken to the next screen: some of your answers to the questions above might be wrong, please review them).

A Pilot Study

[Note: The text below was shown to participants in the Many Avoided condition. The version for the Few Avoided condition differed, with deviations indicated in square brackets.]

The information below gives you an idea of the choices of other people in a similar situation.

We ran a pilot study similar to this experiment, and found that, out of a set of 10 University of Nottingham students, **8 [2] chose the option “Later”**, preferring not to find out the Outcome until the end of Part 2. Therefore, in that group, **the overwhelming majority [minority] chose “Later”**.

If people in your group behave similarly to those in the pilot study, you can expect around 23 [6] people other than you to choose Later, and around 6 [23] to choose Now. In that case, the resulting volume level would roughly be 89 [60] if you choose Now, and 91 [62] if you choose Later.

Out of any 10 members of your group (*other than you*), how many do you believe will choose “Later”?

[Slider ranging from “0 out of 10 will choose “Later”” to “10 out of 10 will choose “Later””]

You are now ready to start the quiz

From the next screen, the quiz will start.

When you feel ready to start, press “Continue”, and both the quiz and the 3-minute timer will start at once.

Quiz

Please, answer the following questions to achieve a higher level and earn a higher payment.

[*List of multiple choice quiz questions is displayed.*]

Your quiz time has ended.

Please, press the Continue button.

The quiz has been interrupted. You will now be asked to make a series of decisions, including your preferred choice between knowing your lottery Outcome “Now” and “Later”.

What do you think the Outcome of your group’s lottery will be? (*Screams* or *Quiet*)
[Radiobuttons (10 levels), with extremes labeled as “Definitely Quiet” and “Definitely Screams”.]

[Note: this screen elicits the Kreps-Porteus questions (KP questions)]

Before choosing between “Now” and “Later”, please answer the following questions about how happy you would be in different (hypothetical) situations.

- Imagine that you choose “Now”, and then a moment later discover that the Outcome is Screams. How happy would you be in that moment *thinking about the Outcome ahead?*

[Slider ranging from -10=“Very unhappy” to 10=“Very happy”.]

- Imagine that you choose “Now”, and then a moment later discover that the Outcome is Quiet. How happy would you be in that moment *thinking about the Outcome ahead?*

[Slider ranging from -10=“Very unhappy” to 10=“Very happy”.]

- Imagine that you choose “Later”, and therefore you do not learn the Outcome until later. How happy would you be in that situation *thinking about the Outcome ahead?*

[Slider ranging from -10=“Very unhappy” to 10=“Very happy”.]

Do you want to know the Outcome “Now” or “Later”?

Choose whether you want to know “Now” or “Later” whether the screams are upcoming.

- ☐ Later
- ☐ Now

How strong is your preference for your selected option, over the other?

[Radiobuttons (10), ranging from “I am **indifferent** between the two options” to “I have a **very strong** preference for the selected option”.]

A quick question

You have chosen the option [“Now”/“Later”]. Why did you prefer this option?

Please take a moment to answer this question.

[Open response textbox.]

Instructions: The Decisions Table

One participant will have the opportunity to **condition their decision on what the other participants chose** in the previous screen. This participant will be selected randomly by the computer.

Suppose you are that randomly selected participant. Then, the decision between “Now” and “Later” that you complete in the table on the following screen will be implemented for you.

See the image below for an example. In the table, you have the possible choices that other participants have made in the previous screen. Each row represents a different situation. **For each row, you can decide which option you would prefer.** Then, according to what the other players have *actually* chosen in the previous screen, the relevant row of the table will be determined, and the option selected in that row of the table will actually be implemented for you.

You have a positive chance of being selected as this player. Therefore, you must complete the table carefully according to your preferences.

(Please scroll down to see the 'Continue' button)

Out of 10 other participants, those who chose <i>Later</i> are...	In that situation, you prefer to know the Outcome...	Volume of screams if you choose Now	Volume of screams if you choose Later
0	Now Later	50	52
2	Now Later	60	62
4	Now Later	69	71
6	Now Later	79	81
8	Now Later	89	91
10	Now Later	98	100

Please, select your preferred option for each of the following situations.

Out of 10 other participants, those who chose <i>Later</i> are...	In that situation, you prefer to know the Outcome...	Volume of screams if you choose Now	Volume of screams if you choose Later
0	Now Later	50	52
2	Now Later	60	62
4	Now Later	69	71
6	Now Later	79	81
8	Now Later	89	91
10	Now Later	98	100

You will now be asked a series of questions. To answer them, please imagine you are in each of the situations presented.

Questions (1/5)

[Note: This and following screens elicit the Information Preferences Scale (IPS; Ho et al., 2021)]

Please, imagine you are in each of the following situations, and answer the questions:

1. As part of a medical checkup, your doctor asks you a series of questions. The answers to these questions can be used to estimate your life expectancy (the age you are predicted to live to). Do you want to know how long you can expect to live?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
2. You provide some genetic material to a testing service to learn more about your ancestors. You are then told that the same test can, at no additional cost, tell you whether you have an elevated risk of developing Alzheimer's (a brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks). Do you want to know whether you have a high risk of developing Alzheimer's?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
3. At your medical checkup, you are given the option to see the results of a diagnostic test, which can identify the extent to which your body has suffered long-term effects from stress. Do you want to know how much lasting damage your body has suffered from stress?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know

- ☐ Probably want to know
- ☐ Definitely want to know

Questions (2/5)

4. Ten years ago, you had the opportunity to invest in two retirement funds: Fund A and Fund B. For the past 10 years, you have invested all your retirement savings in Fund A. Do you want to know the balance you would have if you had invested in Fund B instead?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
5. You decide to go to the theater for your birthday and give your close friend (or partner) your credit card so they can purchase tickets for the two of you, which they do. You aren't sure but suspect that the tickets may have been expensive. Do you want to know how much the tickets cost?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
6. You bought an electronic device at a store at what seemed like a reasonable, though not particularly low, price. A month has passed, and the item is no longer returnable. You see the same device displayed in another store with a sign announcing "SALE." Do you want to know the price you could have bought it for?
 - ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know

Questions (3/5)

7. You gave a close friend one of your favorite books for her birthday. Visiting her apartment a couple of months later, you notice the book on her shelf. She never said anything about it; do you want to know if she liked the book?

- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
8. Someone has described you as quirky, which could be interpreted in a positive or negative sense. Do you want to know which interpretation he intended?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
9. You gave a toast at your best friend's wedding. Your best friend says you did a good job, but you aren't sure if he or she meant it. Later, you overhear people discussing the toasts. Do you want to know what people really thought of your toast?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
10. As part of a fundraising event, you agree to post a picture of yourself and have people guess your age (the closer they get, the more they win). At the end of the event, you have the option to see people's guesses. Do you want to learn how old people guessed that you are?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
11. You have just participated in a psychological study in which all of the participants rate others' attractiveness. The experimenter gives you an option to see the results for how people rated you. Do you want to know how attractive other people think you are?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know

12. Some people seek out information even when it might be painful. Others avoid getting information that they suspect might be painful, even if it could be useful. How would you describe yourself?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
13. If people know bad things about my life that I don't know, I would prefer not to be told.
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know

Questions (5/5)

The questions below ask about what you would decide in the workplace. If you have previously worked in a company, imagine yourself in it and answer the questions. If you had not worked for a company, read the following questions as imagining you are working in a company.

14. Employers commonly encourage workers to complete health risk assessments that gauge their health characteristics and risk. These are usually used to estimate workplace-wide risks the company may face. If you had the opportunity, would you like to get your individual risk assessment?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know
 - ☐ Definitely want to know
15. Most people spend some time at work on activities they are not proud of (e.g., browsing social media). Suppose you could compose a list of these activities and have your computer track how much time you spend doing them. The information would not be shared with your employer. Would you like to know what fraction of each hour at work you spend on these activities?
- ☐ Definitely don't want to know
 - ☐ Probably don't want to know
 - ☐ Probably want to know

- ☐ Definitely want to know
16. As part of a unit-wide evaluation, you and your coworkers are asked how easy it is to get along with each other and what your strengths and weaknesses are. Would you like to know how your coworkers rated you?
- ☐ Definitely don't want to know
- ☐ Probably don't want to know
- ☐ Probably want to know
- ☐ Definitely want to know
17. The rise of artificial intelligence is likely to lead to job losses in a wide range of occupations, affecting workers in industries from long-distance trucking to healthcare. Would you like to know whether your job is at risk owing to automation in the next 10 years?
- ☐ Definitely don't want to know
- ☐ Probably don't want to know
- ☐ Probably want to know
- ☐ Definitely want to know
18. One option in your employer's retirement plan allows you to compare your investment return with that of your coworkers. Would you like to know how your investment return compares with the average and highest investment returns earned by employees at your firm?
- ☐ Definitely don't want to know
- ☐ Probably don't want to know
- ☐ Probably want to know
- ☐ Definitely want to know

Please answer the following questions about how happy you would be in different (hypothetical) situations.

- Imagine that the Outcome is *Screams*, and that the **volume of the screams** will be **level 50**. How happy would you thinking about the Outcome ahead?
[Slider from -10="Very unhappy" to 10="Very happy"]
 - Imagine that the Outcome is *Screams*, and that the **volume of the screams** will be **level 100**. How happy would you thinking about the Outcome ahead?
[Slider from -10="Very unhappy" to 10="Very happy"]
-

Thank you for answering the questions.

In the next screen, you will get to know the Outcome if you selected option Now, or remain without knowing it, if you selected option Later. If you were selected as the random player, then the option you selected in the relevant row of the table will be implemented.

*[Note: this screen was displayed to participants whose implemented option was **Now**]*

Based on your previous decisions, the choice implemented for you is “Now”.

The Outcome is **Screams** [Quiet].

Therefore, you **will hear the screams** [not hear any screams] during the Risk Period.

Please, indicate below which is the Outcome.

- ☐ Screams
- ☐ Quiet

*[Note: this screen was displayed to participants whose implemented option was **Later**]*

Based on your previous decisions, the choice implemented for you is “Later”.

You may or may not hear the screams during the Risk Period. You will be informed about the Outcome later, just before the Risk Period starts.

Press Continue to proceed to Quiz Part 2.

Quiz

Please, answer the following questions to achieve a higher level and earn a higher payment.

[The same list of multiple choice quiz questions is displayed.]

The Quiz has ended.

You will be asked a series of questions, and then move to the Risk Period.

Questions (1/5)

[Note: This and following screens elicit social and economic preferences from the Global Preferences Survey (GPS; Falk et al., 2018)]

We now ask for your **willingness to act** in a certain way in four different areas.

Please indicate your answer on a scale from 0 to 10.

How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future? [Slider: 0=“Completely unwilling to do so”, 10=“Very willing to do so”.]

In general, how willing or unwilling are you to take risks? [Slider: 0=“Completely unwilling to take risks”, 10=“Very willing to take risks”.]

Questions (2/5)

19. When someone does me a favor I am willing to return it.

[Slider: 0=“Does not describe me at all”, 10=“Describes me perfectly”.]

20. Please think about what you would do in the following situation:

You are in an area you are not familiar with, and you realize you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about £20 in total. However, the stranger says he or she does not want any money from you. You have six presents with you. The cheapest present costs £5, the most expensive one costs £30. Do you give one of the presents to the stranger as a “thank-you” gift? If so, which present do you give to the stranger?

- ☐ No present
- ☐ Present worth £5
- ☐ Present worth £10
- ☐ Present worth £15
- ☐ Present worth £20
- ☐ Present worth £25

- ☐ Present worth £30

Questions (3/5)

If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so. [Slider: 0=“Does not describe me at all”, 10=“Describes me perfectly”.]

How willing are you to punish someone who treats others unfairly, even if there may be costs for you? [Slider: 0=“Completely unwilling to do so”, 10=“Very willing to do so”.]

Questions (4/5)

Imagine the following situation: Today you unexpectedly received £1,000. How much of this amount would you donate to a good cause? [Open text box; inputs restricted to numbers between 0 and 1000]

How willing are you to give to good causes without expecting anything in return? [Slider: 0=“Completely unwilling to do so”, 10=“Very willing to do so”.]

Questions (5/5)

“I assume that people have only the best intentions.” [Slider: 0=“Does not describe me at all”, 10=“Describes me perfectly”.]

Instructions: the Risk Period (1/2)

Thanks for kindly answering the questions.

You are now about to enter the Risk Period.

Before the Risk Period starts, two things will happen:

21. In the next screen, you will be provided with a sample scream, to get an idea of the kind of screams that may happen during the Risk Period.

22. After that, everybody will be informed about: a) which is the Outcome (*Screams* or *Quiet*); and b) the volume of the screams as the result of the group's choices, if the Outcome is *Screams*.

After that, the Risk Period will start. It will last 4 minutes, and you will be provided with the answers to the quiz to review if you want to.

Pre-experiencing a sample scream (1/2)

To give you an idea of what the screams sound like, on the next screen, you will hear a sample scream, similar to the screams you would hear if your lottery outcome were *Screams*.

On the next screen, you will hear only 1 scream, whereas during the Risk Period you may hear more than 1 scream if your lottery Outcome is *Screams*, and 0 if your lottery outcome is *Quiet*.

To ensure your safety, the volumes of the screams have been calibrated to comply with the safe hearing levels recommended in the UK.

Please, keep your headphones on at all times. You are also reminded you may leave the session at any time you would like.

Please, try to remain silent throughout.

Now, if you wish to proceed, press Continue.

Pre-experiencing a sample scream (2/2)

To hear the sample scream, press the “Play” button. From the moment you press the “Play” button below, the sample scream will be played at a random moment within 30 seconds.

You will be able to hear the sample scream only once. In other words, the button will not do anything from the second time you press it.

Please, try to remain silent throughout.

After the 30 seconds, the Continue button will appear. You will be able to proceed by pressing it.

[Play button]

The Outcome, revealed

[If the lottery Outcome was Screams, the following text was displayed:]

Your group's lottery Outcome is **Screams**.

Therefore, your group **will hear the screams** during the Risk Period.

The volume level of the screams is [**volume level**]. This is the result of the choices of all participants.

[If the lottery Outcome was Quiet, the following text was displayed:]

Your lottery outcome is **Quiet**.

Therefore, you will **not hear any screams** during the Risk Period.

Risk Period

The Risk Period has started. Please, remain in silence until it ends.

In the meantime, you can read the list of quiz questions and answers below.

At the end of the Risk Period, a “Continue” button will appear, at the bottom. (You may need to scroll down)

Quiz Answers

Below is the list of quiz questions. The correct answers to these are underlined.

[List of all quiz questions and their answers is displayed]

Thank you for your patience. The Risk Period has ended.

You will now be asked a few final questions.

Please, press the Continue button to proceed.

Final Questions (1/4)

Please, kindly answer the following questions.

- How easy to understand were the instructions? [Radiobuttons (5): “Very easy” to “Very hard”]
- How hard to answer were the quiz questions? [Radiobuttons (5): “Very easy” to “Very hard”]
- Do you have any experience at work, even casually? [“Yes”/“No”]

Final Questions (2/4)

- All in all, how do you feel about this study? [Open text box]
- Is there anything else you would want to tell us? [Open text box]

Final Questions (4/4)

- What is your age? [Open text box, restricted to numbers]
- What is your gender? [Female/Male/Other]

You have answered [Q] questions correctly. Therefore, you have reached performance level [P]. In consequence, you have earned £[e] from the quiz.

Adding the £5 for your participation, your **total earnings** from the experiment are £[e + 5].

You will be contacted about your payment in the next couple of days.

Thanks for your participation

We appreciate that you took the time to come and participate in today’s experiment. Thanks a lot, and hope to see you around soon!

Your session is ending now. You can now leave the lab. Please, do it in silence, so other participants are not disturbed.

B Appendix for Chapter 2

B.1 Dynamics of confirmation-biased learning

To get a better feel of the implications of the confirmation-biased learning process, it is useful to examine the evolution of beliefs over time. We begin at $t = 1$, when the very first piece of raw information y_1 arrives. The information is first interpreted, as follows:

$$\begin{aligned}\hat{y}_2 &= \mu_1 \left(\frac{h_1}{h_2^y} \right) + y_1 \left(\frac{h_\epsilon}{h_2^y} \right), \\ h_2^y &= h_1 + h_\epsilon.\end{aligned}$$

In the confirmatory biased interpretation $\hat{y}_2$, the location of the information y_1 is pulled towards the initial prior mean μ_1 . In addition, by conforming the information to the initial belief, the variance of the information is reduced, and its perceived precision h_2^y is increased by the addition of one extra prior precision h_1 . In other words, a confirmatory biased evaluator overestimates the precision of the interpreted information.

The evaluator then updates her beliefs based on her interpretation of the information $\hat{y}_2$:

$$\begin{aligned}\hat{\mu}_2 &= \mu_1 \left(\frac{h_1}{h_2^\mu} \right) + \hat{y}_2 \left(\frac{h_2^y}{h_2^\mu} \right), \\ h_2^\mu &= h_1 + h_2^y = 2h_1 + h_\epsilon.\end{aligned}$$

The interpretive learner ends up with a posterior precision of $2h_1 + h_\epsilon$, whereas a Bayesian evaluator would have a precision of only $h_1 + h_\epsilon$. Hence, an interpretive evaluator features overconfidence about her beliefs. This is one of the features of confirmation bias, as Rabin and Schrag (1999) have noted and demonstrated as well in a discrete model.

We can express the interpretive posterior mean as a function of the raw information:

$$\hat{y}_2 = \mu_1 \left(\frac{2h_1}{2h_1 + h_\epsilon} \right) + z_1 \left(\frac{h_\epsilon}{2h_1 + h_\epsilon} \right).$$

The prior belief μ_1 is weighted twice, more than a Bayesian would, which has the effect of pulling the posterior even closer to the initial belief μ_1 . In effect, confirmation bias produces an overweight of initial beliefs, which ultimately leads to prior inertia in the belief-updating process.

At $t = 2$, the evaluator gets a new, independent piece of information. A Bayesian evaluator would treat this information as independent. But the interpretation step is as follows:

$$\begin{aligned}\hat{y}_3 &= \hat{\mu}_2 \left(\frac{h_2^\mu}{h_3^y} \right) + y_2 \left(\frac{h_\epsilon}{h_3^y} \right), \\ h_3^y &= h_2^\mu + h_\epsilon = 2h_1 + 2h_\epsilon, \\ \hat{y}_3 &= \mu_1 \left(\frac{2h_1}{2h_1 + 2h_\epsilon} \right) + y_1 \left(\frac{h_\epsilon}{2h_1 + 2h_\epsilon} \right) + y_2 \left(\frac{h_\epsilon}{2h_1 + 2h_\epsilon} \right).\end{aligned}$$

At $t = 2$, the precision h_3^y of the interpretation $\hat{y}_3$ is $2h_1 + 2h_\epsilon$, higher than the precision of the raw information y_2 (which is only h_ϵ). Notice how the initial prior precision h_1 is added twice in the interpretation of the second period information (more than what happened with the interpretation at $t = 1$). Moreover, the precision of previous information y_1 is added once in the interpretation of this *independent* second piece of information, via the prior $\hat{\mu}_2$.

Now, looking at the mean of the interpreted information, $\hat{y}_3$, it is not only the case that the initial belief μ_1 is overweighted again (this time, appearing twice). It is also the case that the first piece of information y_1 has an influence on the interpretation of the second, independent piece of information. Early information thus influences the interpretation of subsequent information. This is consistent with the classical Asch (1946) experiments, which empirically showed that early information conditions the interpretation of later information.

Mathematically, the influence of early signals on the interpretation of subsequent signals is shown by the presence of y_1 in the interpretation $\hat{y}_3$. Based on this interpretation, the posterior belief is updated as:

$$\begin{aligned}\hat{\mu}_3 &= \hat{\mu}_2 \left(\frac{h_2^\mu}{h_3^\mu} \right) + \hat{y}_3 \left(\frac{h_3^y}{h_3^\mu} \right), \\ h_3^\mu &= h_2^\mu + h_3^y = 4h_1 + 3h_\epsilon, \\ \hat{\mu}_3 &= \mu_1 \left(\frac{4h_1}{4h_1 + 3h_\epsilon} \right) + y_1 \left(\frac{2h_\epsilon}{4h_1 + 3h_\epsilon} \right) + y_2 \left(\frac{h_\epsilon}{4h_1 + 3h_\epsilon} \right).\end{aligned}$$

By $t = 2$ the initial prior gets 4 times its precision. Moreover, notice that the early signal y_1 is now given twice the precision that a Bayesian evaluator would assign. This corresponds to another effect of confirmation bias, a preference for early information, which is a corollary of the fact that early information influences the interpretation of future information. Notice how the information is weighted more strongly when it appears earlier in the series. Because of this, confirmation-biased evaluators are more influenced by first impressions. This implies that the exchangeability property of Bayesian learning does not hold anymore, and confirmation-biased learning features informational path-dependencies instead.

To grasp what happens as information grows, consider $t = 4$. By $t = 4$, the interpretation rule becomes:

$$\begin{aligned}\hat{y}_5 &= \hat{\mu}_4 \left(\frac{h_4^\mu}{h_5^y} \right) + y_5 \left(\frac{h_\epsilon}{h_5^y} \right), \\ h_5^y &= h_4^\mu + h_\epsilon = 8h_1 + 8h_\epsilon, \\ \hat{y}_3 &= \mu_1 \left(\frac{8h_1}{8h_1 + 8h_\epsilon} \right) + y_1 \left(\frac{4h_\epsilon}{8h_1 + 8h_\epsilon} \right) + y_2 \left(\frac{2h_\epsilon}{8h_1 + 8h_\epsilon} \right) \\ &\quad + z_3 \left(\frac{h_\epsilon}{8h_1 + 8h_\epsilon} \right) + z_4 \left(\frac{h_\epsilon}{8h_1 + 8h_\epsilon} \right),\end{aligned}$$

based on which the posterior belief is updated as:

$$\begin{aligned}\hat{\mu}_5 &= \hat{\mu}_4 \left(\frac{h_4^\mu}{h_5^\mu} \right) + \hat{y}_5 \left(\frac{h_5^y}{h_5^\mu} \right), \\ h_5^\mu &= h_4^\mu + h_5^y = 16h_1 + 15h_\epsilon, \\ \hat{\mu}_5 &= \mu_1 \left(\frac{16h_1}{16h_1 + 15h_\epsilon} \right) + y_1 \left(\frac{8h_\epsilon}{16h_1 + 15h_\epsilon} \right) + y_2 \left(\frac{4h_\epsilon}{16h_1 + 15h_\epsilon} \right) \\ &\quad + y_3 \left(\frac{2h_\epsilon}{16h_1 + 15h_\epsilon} \right) + y_4 \left(\frac{h_\epsilon}{16h_1 + 15h_\epsilon} \right).\end{aligned}$$

By recursion, the rules can be expressed as a function of t as:

$$\begin{aligned}\hat{\mu}_{t+1} &= \mu_1 \left(\frac{h_1}{h_1 + h_\epsilon} \right) + \left[\sum_{j=1}^{t-1} y_j \frac{1}{2^j} + y_t \frac{1}{2^{t-1}} \right] \left(\frac{h_\epsilon}{h_1 + h_\epsilon} \right) \\ h_{t+1}^y &= 2^{t-1} (h_1 + h_\epsilon) \\ \hat{\mu}_{t+1} &= \mu_1 \left(\frac{2^t h_1}{2^t (h_1 + h_\epsilon) - h_\epsilon} \right) + \sum_{j=1}^t y_j \left(\frac{2^{t-j} h_\epsilon}{2^t (h_1 + h_\epsilon) - h_\epsilon} \right) \\ h_{t+1}^\mu &= 2^t (h_1 + h_\epsilon) - h_\epsilon\end{aligned}$$